Segmentation des Traits du Visage, Analyse et Reconnaissance des Expressions Faciales par les Modèles de Croyance Transférable.

Zakia Hammal

UNIVERSITE JOSEPH FOURIER DE GRENOBLE

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UJF

Spécialité : "Sciences Cognitives"

préparée au Laboratoire des Images et des Signaux
dans le cadre de ***l'École Doctorale "Ingénierie pour la Santé, la Cognition et l'Environnement"***

présentée et soutenue publiquement

par

Zakia Hammal

le 29 Juin 2006

Titre :

Segmentation des Traits du Visage, Analyse et Reconnaissance d'Expressions Faciales par le Modèle de Croyance Transférable

Facial Features Segmentation, Analysis and Recognition of Facial Expressions using the Transferable Belief Model

---

Directrice de thèse : ALICE CAPLIER

JURY

MR. Jeanny HERAULT,	Président
MR. Patrice DALLE,	Rapporteur
MR. Thierry DENOEUX,	Rapporteur
MME. Michèle ROMBAUT,	Examinatrice
MR. Thierry PUN,	Examinateur
MME. Alice CAPLIER,	Directrice de thèse

A mes parents, ma soeur et mes deux frères

Remerciements

Tout d'abord, je remercie mon encadrante Alice pour la pertinence de ses conseils et sa rigueur scientifique.
Je remercie également les membres du jury:
- Jeanny Hérault, pour avoir présidé la soutenance.
- Les rapporteurs, Thierry Denoeux et Patrice Dalle, pour la finesse et l'exactitude de leurs analyses. Les questions qu'ils ont soulevées m'ont donné matière à réflexion pour encore quelques temps.
- Les examinateurs, Michèle Rombaut et Thierry Pun, pour avoir examiné ma thèse.

Des remerciements plus particuliers pour Michèle Rombaut qui a aussi été présente pendant les deux dernières années de ma thèse. Je la remercie pour sa rigueur, ses conseils tout au long de nos nombreuses discussions et surtout pour sa richesse humaine. Je la remercie également pour les pistes qu'elle m'a permis d'explorer et surtout pour son soutien tout au long de ma thèse et de ma rédaction.

Je remercie également mes parents pour leur soutien, leurs encouragements et leur confiance sans faille depuis le début.

Mes remerciements vont aussi à mes ami(es) du LIS. Tout d'abord mes compagnons de thèse Corentin, Pierre et Mickael pour tout ce qu'ils m'ont apporté au laboratoire et en dehors. Plus particulièrement mon collègue de bureau Corentin pour son soutien pendant toutes ces années. Une pensée particulière aussi à Hervé pour son amitié pendant et après son départ du LIS. Alan pour son amitié et ses conseils.

Je remercie également Nadia pour sa douceur et sa gentillesse qui m'ont souvent redonné le sourire. Un grand merci à Jeanny pour m'avoir permis de m'inscrire en thèse, GG (Bouvier) pour sa bonne humeur et sa gentillesse depuis mon arrivée au LIS, Michel (Crepin-Jourdan) et Marino pour leur gentillesse. Danuta et sa famille pour la chaleur familiale qu'ils m'ont offerte.

Un merci aussi à mes amis en transit au LIS, plus particulièrement Guillermo de Barcelone et maintenant au Pays-Bas pour son amitié et sa richesse humaine, et surtout pour m'avoir fait découvrir l'Espagne et plus particulièrement Barcelone. Merci aussi à son épouse Teresa pour sa gentillesse. Une pensée aussi à Jordi d'Espagne mon compagnon des soirs et des week-end au LIS dans les périodes les plus vides du labo et pour ses messages de soutien.

Merci à mes amis de Paris, Vincent et Aurélien, et surtout Vincent pour son soutien sans faille et ses conseils tout au long de sa présence au LIS et pour son amitié depuis son départ.

Un grand merci aussi à mes ami(es) et collègues de Belgique. Tout d'abord Thierry Dutoit pour m'avoir accueillie dans son laboratoire et m'avoir intégrée dans son équipe. Je remercie Baris, Laurent, Devrim et Céline pour leur accueil et leur amitié pendant mon séjour en Belgique et pour tout ce qu'ils m'ont apporté au niveau scientifique. Je les remercie également pour tous les messages de soutien et d'encouragement pendant ma rédaction. Je remercie aussi mon compagnon du soir à Multitel Olivier d'avoir souvent été à ma disposition pour me raccompagner les soirées tardives et les nuits blanches à Mons en Belgique.

Un merci aussi à Stéphane, Denis (Pellerin), Jocelyn, Denis, Frédéric, Mériam, Vincent, Sebastian, Anthony, Bertrand, Cedric, Sébastien et Barbara.

Un grand merci aussi à mes amis de Grenoble, Florence et Jean François pour avoir été une famille d'acceuil et pour m'avoir conseillée et soutenue depuis mon arrivée à Grenoble. Une pensée aussi à Jérome, Chloé et Amandine. Aux copines des cours de danse de Saint-Pierre pour m'avoir intégrée dans leur groupe et m'avoir permis de m'évader au travers d'un spectacle de danse. Mes ami(es) de Caen Salah et surtout Aurore pour m'avoir soutenue pendant le DEA et la Thèse.

Une thèse c'est beaucoup de savoir, de savoir faire mais c'est surtout une aventure humaine très enrichissante par ses bons et ses ... côtés. Je suis contente de l'avoir vécue avec vous tous.

Contents

List of Figures

List of Tables

Introduction

Ces dernières années, les ordinateurs sont devenus de plus en plus présent dans la vie quotidienne de la population. Impliqués aussi bien dans les activités professionnelles que personnelles pour réaliser des tâches de plus en plus complexes, leur accessibilité doit être améliorée autant que possible. Ils doivent donc intégrer le contexte et des capacités d'interprétation du comportement de l'utilisateur afin de répondre à ses demandes.

Durant les dix dernières années, les interfaces homme-machine étaient essentiellement basées sur des interactions simples utilisant des modalités classiques (par exemple, clavier, souris, écran tactile, etc..). Dans le but d'atteindre une interaction homme-machine efficace, l'ordinateur doit pouvoir interagir avec l'utilisateur aussi naturellement que possible, de la même manière qu'une interaction humaine.

Ces dernières années ont marqué un intérêt grandissant pour l'amélioration de tous les aspects d'interaction entre l'homme et la machine dans le but de développer des interfaces multimodales intelligentes.

Ce domaine émergent a été le centre d'intérêt des recherches scientifiques de différentes voies scolastiques, telles que la vision par ordinateur, l'ingénierie, la psychologie et les neurosciences. Ces études se sont intéressées non seulement à l'amélioration des interfaces d'ordinateur, mais aussi à l'amélioration des actions que l'ordinateur peut exécuter en se basant sur le retour donné par l'utilisateur. Par exemple dans la recherche sur internet, il peut être utile de savoir si l'utilisateur s'ennuie ou s'il n'est pas satisfait des résultats de la recherche. En apprentissage, il peut être utile d'aider de manière efficace des élèves utilisant l'ordinateur comme professeur quand ils sont face à des problèmes ou quand ils s'ennuient, ou au contraire, quand ils sont contents. A cette fin différentes modalités peuvent être utilisées. Par exemple la voix d'un utilisateur peut être enregistrée par un microphone et interprétée comme du texte (*speech to text*), les mots peuvent être reconnus et la parole synthétisée (*text to speech*); le visage de l'utilisateur peut être capturé par une caméra et son expression faciale identifiée; de même les gestes des mains peuvent être suivis et leurs mouvements interprétés.

Parmi tous, un des ingrédients nécessaire dans une interaction naturelle en face à face est l'expression faciale. Charles Darwin a été l'un des premiers scientifiques à reconnaître que l'expression faciale est le moyen le plus puissant et immédiat chez les êtres humains pour communiquer leurs émotions, intentions, et opinions. De plus une expression faciale est une manifestation visible de l'état émotionnel, de l'activité cognitive, de l'intention, de la personnalité et de la psychopathologie d'une personne [Don99]. Mehrabian [Meh68] a mis en évidence que les 55% d'un message émotionnel sont communiqués par l'expression faciale alors

que 7% seulement par le canal linguistique et 38% par le paralanguage (comme l'intonation). Ceci implique que les expressions faciales jouent un rôle important dans la communication humaine. Donc l'interaction homme-machine tirera un véritable bénéfice d'un système automatique de reconnaissance d'expressions faciales.

Dans le travail présenté nous nous sommes intéressés à la reconnaissance des expressions faciales dans une séquence vidéo. Comme tout autre comportement humain, reconnaître une expression faciale est une tâche complexe à accomplir par un système de vision par ordinateur à cause de la grande variabilité entre les individus. De nombreux travaux ont été proposés pour la reconnaissance et l'interprétation des expressions faciales. Ces travaux ont fait émerger deux questions importantes : quels sont les indices pertinents qui doivent être extraits d'un visage? Comment le comportement de ces indices peut être modélisé et traduit pour la reconnaissance des expressions faciales?

Dans ce travail nous avons essayé d'apporter notre contribution à la réponse à ces deux questions.

Notre but est la reconnaissance des expressions faciales d'un utilisateur face à son écran d'ordinateur dans des conditions de travail normales. L'information en entrée du système est une image numérique du visage de l'utilisateur. La première étape est donc d'identifier qu'elle est l'information présente dans un visage qui va être utilisée pour reconnaître une expression.

Avec le concours du laboratoire de psychologie sociale de Grenoble et du CLIPS, nous avons mis en place une expérience psychologique qui nous a permis de conclure que les contours des yeux, des sourcils et des lèvres, sont des informations qui doivent être prises en compte dans le processus de reconnaissance des expressions faciales. Notre système de reconnaissance des expressions analysera ensuite le comportement de ces traits permanents.

En se basant sur ces observations, nous proposons une méthode de segmentation des traits permanents du visage. L'extraction de ces données bas niveau a conduit à plusieurs travaux avec différentes approches. Notre méthode se focalise sur l'extraction des contours précis, flexibles et robustes aux déformations des traits du visage.

Par la suite, et dans le but de mesurer les déformations de ces traits, nous définissons un ensemble de distances caractéristiques. Nous leur associons un ensemble de règles logiques pour décrire chaque expression faciale. Ces règles sont basées sur la description MPEG-4 des expressions faciales et aussi sur notre propre analyse des déformations des traits du visage sur les expressions étudiées.

Basé sur ces règles logiques, un processus de fusion de toutes les distances caractéristiques est requis pour finalement reconnaître les expressions faciales. Les systèmes de reconnaissance d'expressions existant effectuent une classification des expressions examinées en une des émotions de base proposées par Ekman et Friesen [Ekm78]. Cependant l'être humain n'est pas binaire et des expressions pures sont rarement produites. Basé sur ces observations et dans le but de modéliser le doute entre plusieurs expressions faciales le Modèle de Croyance Transférable (MCT) est utilisé comme processus de fusion pour la classification des expressions faciales.

Cependant dans la vie de tous les jours une expression faciale n'est pas une information statique mais le résultat d'une évolution temporelle des déformations des traits du visage. Bassili [Bas78] a montré qu'une expression faciale est mieux reconnue dans une séquence vidéo que dans une image statique. Sur la base de ces observations nous proposons d'introduire l'information temporelle dans le processus de classification pour une classification dynamique des séquences d'expressions faciales.

Pour allez plus loin, dans une communication humaine les individus arrivent souvent à reconnaître les émotions à partir des expressions faciales perçues et du ton de la voix. Ceci est une information importante participant à la teneur de l'échange linguistique. Comme de plus en plus d'ordinateurs sont équipés avec un matériel audio visuel, il devient concevable que l'ordinateur puisse être entraîné à effectuer la même inférence. Par conséquent nous avons également initié une étude sur l'analyse et la reconnaissance des expressions vocales basées sur l'analyse de caractéristiques prosodiques.

Finalement nous présenterons nos perspectives pour fusionner les deux modalités dans le but d'obtenir un système bimodal de reconnaissance d'expressions utilisant les deux modalités image et son.
Ce mémoire est constitué de trois principales parties. La première se focalise principalement sur l'extraction de données bas niveau, la seconde sur le processus de classification et la troisième sur les travaux préliminaires vers un système bimodal de classification d'expressions.

Dans la Partie I, premièrement l'ensemble de contraintes morphologiques et les bases de données utilisées dans le processus de segmentation sont décrites dans le chapitre 1. Le chapitre 2 propose, d'une part un aperçu général des méthodes existantes et d'autre part une description des algorithmes de segmentation proposés pour la segmentation de l'iris (section 2.2.2), des yeux, des sourcils (section 2.4) et de la bouche (section 2.5). Finalement la section 2.7 propose un ensemble de résultats et de discussions sur les performances et limites des algorithmes proposés.

Dans la Partie II, le chapitre 3 décrit dans un premier temps un aperçu général des méthodes existantes pour la classification des expressions faciales et dans un second temps la présentation de notre méthode de classification basée sur le MCT. Le chapitre 4 décrit la classification dynamique des expressions faciales basée sur une modélisation temporelle du MCT.

Dans la Partie III, un état de l'art sur les méthodes existantes pour l'analyse et la classification des expressions vocales ainsi que les méthodes de classification bimodale sont présentés. Nous décrivons ensuite nos travaux préliminaires pour la classification des expressions vocales. Finalement une discussion sur la combinaison des deux modalités est présentée.

Introduction

In last years, computers become more and more present in daily life of the general population. Involved in professional as well as personal activities to realize tasks more and more complex, their accessibility needs to be improved as much as possible. To do so they need to integrate context and user behavior interpretation abilities in order to answer to its expectations.

During the past decades, Human-Computer Interfaces have been relying on simple interactions through classical devices (e.g., keyboard, mouse, touch-screen, etc). To truly achieve effective human-computer interaction, the computer must be able to interact naturally with the user, similarly to the way a human being interacts with another human being. In recent years there has been a growing interest in improving all aspects of the interaction between humans and computers towards the development of intelligent multi-modal interfaces.

This emerging field has been a research interest for scientists from several different scholastic tracks, such as computer science, engineering, psychology, and neuroscience. These studies focus not only on the improvement of computer interfaces, but also on the improvement of the actions the computer can take based on feedback provided by the user. For example in Web browsing or searching, it would be useful to know if the user is bored or dissatisfied with search results. In education, it would be useful to efficiently help students using the computer as a virtual tutor when they are facing problems or when they are bored, or, on the contrary, when they are pleased. For this purpose different modalities can be employed. For example the user voice can be recorded via a microphone and interpreted to text (*speech to text*), words can be recognized and speech synthesized (*text to speech*); user face can be captured via a camera and his/her facial expression can be identified; likewise user hand gesture can be tracked and its movement can be interpreted.

Among others, one necessary ingredient for natural interaction in face-to-face human communication is facial expression. Charles Darwin was one of the first scientists to recognize that facial expression is one of the most powerful and immediate means for human beings to communicate their emotions, intentions, and opinions to each other. In addition a facial expression is a visible manifestation of the emotional state, cognitive activity, intention, personality, and psychopathology of a person [Don99]. Mehrabian [Meh68] pointed out that 55% of the communicating feelings is conveyed by the facial expression while only 7% by the linguistic language and 38% by the paralanguage (like intonation). This implies that facial expressions play an important role in human communication. Then human-computer interaction will definitively benefit from automated facial expression recognition.

In the presented work we are interested in the automatic recognition of facial expressions in video. As any human behavior, recognizing a facial expression is a complex task to be achieved by a computer vision system due to the great variability between people. Several works have been proposed for the recognition and the interpretation of the facial expressions. Two major questions have emerged: what are the pertinent cues that have to be extracted from the face? How can the behavior of these cues be modeled and traduced for the recognition of facial expressions?

In this work we have tried to bring our contribution to respond to these two questions.

Our aim is the recognition of facial expressions of a user facing his computer screen in usual work conditions. The input information of the system is the digital image of the user's face. The first step is thus to identify what is the information presents in a face image that will be used to recognize an expression.

We have carried out a psychological experiment which has led us to consider that the contours of the eyes, eyebrows and mouth are information that have to be taken into account in the recognition process of facial expressions. Our facial expression recognition system will then analyse the behavior of these permanent facial features.

Based on these observations, we propose a segmentation method of the permanent facial features. The extraction of these low level data have led to numerous works with different approaches. Our method focuses on the extraction of contours being as robust and flexible as possible to the facial features deformations.

Then, in order to measure the deformations of these features, we define a set of characteristic distances. We associate to them a set of logical rules to describe each facial expression. These rules are based on the MPEG-4 description of facial expressions and also on our analysis of the deformations of the facial features during the production of the studied expressions.

Based on these logical rules, a fusion process of all the characteristic distances is required to finally classify the facial expressions. The existing expression analyzers perform a classification of the examined expressions into one of the basic emotion categories proposed by Ekman and Friesen [Ekm78]. However people are not binary and pure facial expressions are rarely produced. Moreover, people are different and most of time show *mixture* of facial expressions. Based on these observations and in order to model a doubt between some facial expressions the Transferable Belief Model (TBM) are applied as a fusion process for the facial expressions classification.

However in daily life a facial expression is not a static information but is the result of a temporal evolution of the facial features deformations. Bassili [Bas78] showed that facial expressions can be more accurately recognized from an image sequence than from a static image. Based on these observations we propose to introduce a temporal information in the classification process for dynamic classification of facial expressions sequences.

To go further, in human-to-human communications, people often infer emotions from perceived facial expressions and voice tones. This is an important addition to the linguistic content of the exchange. As more and more computers are equipped with auditory and visual input devices, it becomes conceivable that computers may be trained to perform similar inference. Therefore we have also initiated the study of vocal expressions analysis and classification based on the analysis of prosodic vocal characteristics.

Finally we will present our perspectives to fuse the two modalities towards a bimodal expressions classification system based on the face and voice modalities.

The general organization of the work is described in three main parts. The first one focuses mainly on low-level data extraction, the second one on the classification process and the third one on preliminary work towards bimodal expression classification.

In Part I, firstly a set of morphological constraints and databases used for the segmentation process are described in Chapter 1. The description of the databases we used is also added in this chapter. Chapter 2 proposes first a general overview of the existing methods; then the facial features segmentation algorithms for iris (section 2.2.2), eyes and eyebrows (section 2.4) and mouth (section 2.5) are presented. Finally section 2.7 proposes a set of segmentation results and discussions about the performances and limits of the proposed algorithms.

In Part II chapter 3 describes first a general overview of the existing methods about facial expressions classification. Then we present our static classification method based on the TBM. Chapter 4 describes the dynamic facial expressions classification based on the temporal modeling of the TBM.

In Part III, firstly a state of the art on the existing methods on vocal expression analysis and classification are presented as well as existing bimodal methods. We describe our preliminary work on vocal expressions classification. Finally a discussion on the combination of the two modalities is presented.

Part I

Facial features segmentation: low level data extraction

Préambule. Le problème de reconnaissance des expressions faciales a conduit ces vingt dernières années à de nombreux travaux. L'analyse des expressions d'un visage humain nécessite des pré-traitements qui consistent à détecter ou à suivre le visage, détecter les traits caractéristiques du visage tels que les yeux, les sourcils et la bouche et finalement à extraire et à suivre leur mouvement. Ces mouvements induisent des déformations des traits du visage traduits par exemple par le mouvement des points caractéristiques du visage ou par des modèles de gestes faciaux basés sur des contraintes morphologiques du visage.

Le but de notre travail est un système automatique de classification d'expressions faciales basé sur l'analyse des déformations des traits permanents du visage. La première étape de ce système consiste à segmenter les traits permanents du visage tels que les yeux, les sourcils et la bouche. Dans cette partie nous présentons notre contribution pour la segmentation de ces traits faciaux dans le but d'obtenir des contours précis, flexibles et robustes aux expressions faciales.

Preamble. The problem of recognizing facial expressions has led to numerous works in the last twenty years. Analysis of expression of human face requires a number of preprocessing steps which attempt to detect or track the face, to detect the characteristic facial features such as eyes, eyebrows and mouth and to extract and follow the movements of these facial features. These movements induces deformations of facial features such as, for example, characteristic facial points or model of facial gesture using some morphological constraints about the face.

The aim of our work is an automatic facial expressions classification system based on the analysis of the permanent facial features deformations. Then the first step of this system consists in segmenting the permanent facial features from face image. Various techniques have already been suggested to extract the permanent facial features such as eyes, eyebrows and mouth. In this part we present our contribution for the segmentation of these facial features aiming at obtaining accurate and flexible contours robust to facial deformations.

Chapter 1

Preliminary study and morphological constraints

Facial features are objects easily identifiable by the human visual system. However, developing an automatic system for their detection and localization is not an easy task. Moreover the robustness and the accuracy of this detection process is critical for the usefulness of such system. The methods proposed here are dedicated to iris, eyes and eyebrows segmentation. Before describing our segmentation techniques a preliminary study on a set of databases is required to extract basic knowledge on face properties and facial feature organization.

1.1 Databases

We present here all the databases used to validate all the algorithms presented in this work.

1.1.1 ORL database [Orl94]

The ORL Face (ORLF) database contains a set of face images. The database has been acquired for the aim of face recognition. There are ten different images of 40 distinct subjects. For some subjects, the images were taken with different lighting conditions, facial expressions (open, closed eyes, smiling, not smiling) and facial details (glasses, no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). The files are in PGM format. The size of each image is 92*112 pixels, with 256 gray levels per pixel (Figure 1.1).

1.1.2 Hammal database [Ham03]

The Hammal Face (HF) database contains a set of sequences acquired with a digital camera at 25 frames per second. It is made of 6 different sequences. The subjects were asked to sit down in front of the computer screen, with roll and pan head motion, and with different gaze directions (looking forwards, at the right, at the left, at the bottom, and at the top) to have different iris positions. The sequences contain from 120 to 600 frames. All frames are in color in RGB format. The size of each image is $280 * 380$ pixels (Figure 1.2).

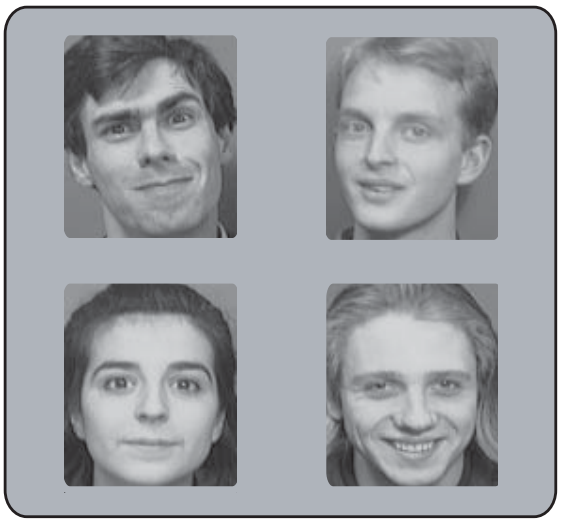

Figure 1.1: Examples of faces from ORL database [Orl94].

Figure 1.2: Examples of frames from HF database [Ham03].

1.1.3 Ferret database [Fer03] and Yale database [Yal97]

The Feret Face (FF) database has been acquired for the purpose of face recognition. There are 994 different subjects and more than 3500 color images in PPM format. The size of each image is equal to $780 * 520$ pixels (Figure 1.3).

The Yale Face (YF) database has been acquired for the aim of face recognition too. There are 15 different subjects with 11 images per subject, one per different facial expression or configuration: left-light, right-light, center-light, with or without glasses, happy, normal,

sad, sleepy, surprised, and wink. There are a total of 165 gray level images in GIF format. The size of each image is equal to $280 * 320$ pixels (Figure 1.5).

Figure 1.3: Examples of frames from FF database [Fer03].

1.1.4 Hammal-Caplier database [Ham03]

The Hammal-Caplier Expressions (HCE) database has been acquired with a digital camera at 25 frames per second. The database has been acquired for the aim of facial expressions recognition. There are 21 different subjects with 3 sequences per subject (*Smile*, *Surprise* and *Disgust* sequence). Each sequence is recorded during 5 seconds. For each different acquisition the subject is asked to simulate one expression beginning by the *Neutral* expression, evolving to the considered expression and coming back to the *Neutral* expression (which gives images with several expressions intensities). There is no constraint on the lighting conditions (indoor acquisition). The database is made of 2520 color frames in BMP format. The size of the acquired faces varies from 90x90 to 200x180 pixels (Figure 1.4).

1.1.5 Dailey-Cottrell database [Dai01] and CKE database [Coh00]

The Dailey-Cottrell Expressions (DCE) database has been acquired for the aim of facial expressions recognition. There are 16 different subjects (8 females and 8 males). For each subject, the images have been acquired under six expressions (*Smile*, *Surprise*, *Disgust*, *Fear*, *Sadness* and *Anger*). For each expression there are two images: an expressive face and a neutral face. The database is then composed of 192 gray level frames in BMP format. The size of the acquired faces is 240x380 pixels (Figure 1.5).

The Cohn-Kanade Expressions (CKE) database [Coh00] has been acquired for the aim of facial expressions recognition. There are 100 different subjects. 65 were females, 15 were African-American, 3 were Asian or Latino. Subjects were instructed by an experimenter to perform six facial expressions (*Smile*, *Surprise*, *Disgust*, *Fear*, *Sadness* and *Anger*). Image sequences from *Neutral* to target expression were digitized into 640x480 gray level frames in BMP format (Figure 1.5).

Figure 1.4: Examples of frames from the HCE database [Ham03].

Figure 1.5: Examples of frames from Top: CKE (a and b) [Coh00] and YF database (c and d) [Yal97]; bottom : DCE database (e and f) [Dai01].

1.2 Morphological constraints

In order to reduce the complexity of the facial features segmentation problem, the first step consists in taking into account some morphological constraints of the facial features: dimensions and relative positions of iris, eyes and eyebrows in static images as well as during dynamic

deformations on different video sequences. We extract a set of geometrical constraints which are introduced in the segmentation process.

Among several face databases the ORLF (see section 1.1.1) database is chosen to realize the statistical analysis on the morphological constraints of eyes and eyebrows bounding boxes. The reason is that each subject has been recorded with systematically different head positions. Then it is possible to study these morphological constraints and to validate them for different positions on 40 subjects.

Statistical analysis of the distance between iris and eyes corners constraints requires sequences with dynamic iris motion under different positions. HF (see section 1.1.2) database is then used because in all the other databases people are always looking forward.

1.2.1 Dimensions and positions of eyes and eyebrows bounding boxes

On each frame of the ORLF database (see section 1.1.1), we manually select a face bounding box, two eyes bounding boxes and two eyebrows bounding boxes. The selected rectangle of the face bounding box passes closely to the ears for its right-hand side limit and left-hand side limit. It is limited at the bottom by the chin and at the top by the half of the forehead. The eyes bounding boxes are limited by the face limits in the right-hand side and left-hand side, by the lower eyelids (a bit bellow) at the bottom and at the top a bit upon the eyebrows (Figure 1.6.left). Finally, the eyebrows bounding boxes are delimited at the bottom by the iris and at the top, right and left by the face bounding boxes (Figure 1.6.right). We measure the size of all selected boxes (Figure 1.6) relatively to the corresponding size of the face box. The obtained mean relations are:

$$H_{eye} = \frac{1}{3} * H_{face} \tag{1.1}$$

$$W_{eye} = \frac{2}{5} * W_{face} \tag{1.2}$$

$$H_{eyebrows} = \frac{1}{4} * H_{face} \tag{1.3}$$

$$W_{eyebrows} = \frac{1}{2} * W_{face} \tag{1.4}$$

with (W_{face},H_{face}), the face bounding boxes dimensions and (W_{eye},H_{eye}), the eyes bounding boxes dimensions and ($W_{eyebrows}$,$H_{eyebrows}$) the eyebrows bounding box dimensions.

1.2.2 Iris position in the face

The iris position in the face has been defined using all the databases (see section 1.1). As a result, until a roll $\approx 45^o$ and a pan $\approx 40^o$ head motion, the following constraints are always verified:

- Irises are always in the higher half of the face.
- The right iris is in the right half of the face.
- The left iris is in the left half of the face.

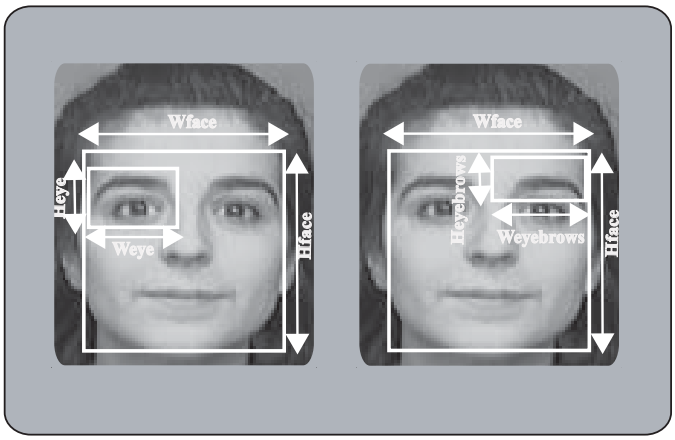

Figure 1.6: Size constraints of eyes and eyebrows bounding boxes relatively to the face dimensions.

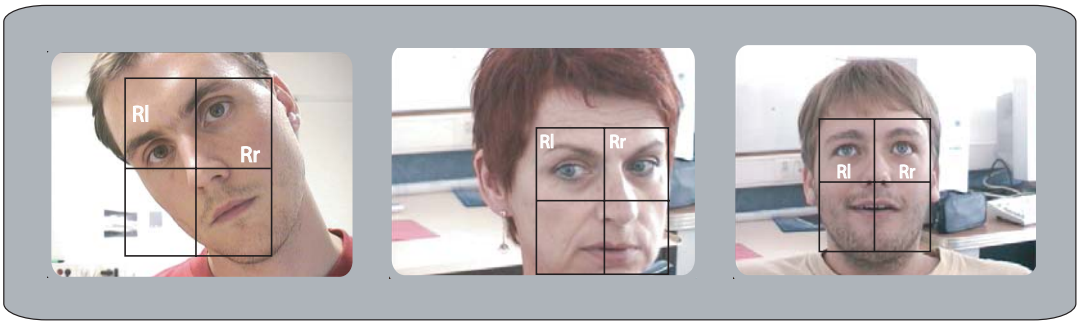

Figure 1.7: Selection of the search zones of the irises; Rl: region of the left iris, Rr: region of the right iris

1.2.3 Distance between irises centers

In order to analyze the distance between the irises centers in different gaze directions, the training is carried out on the HF database (see section 1.1.2). As a result, the distance between the centers of the irises is stable when irises are moving in a frontal view of the face (Figure.1.8). In case of maximum authorized head pan motion, this distance decreases of a maximum amount of $3/2 * R$ (R being the radius of the iris).

1.2.4 Eyes corners positions relatively to iris centers positions

The training is carried out on the HF database (Figure 1.9). As a result, it is possible to roughly localize their horizontal and vertical positions relatively to the center of each iris. The morphological constraints extracted from this training step are:

- The vertical position of the two eyes corners (interior and external corner) is always lower than the iris center and approximatively at $R/2$ pixels below the center of the iris (R being the radius of the iris) (Figure.1.9).

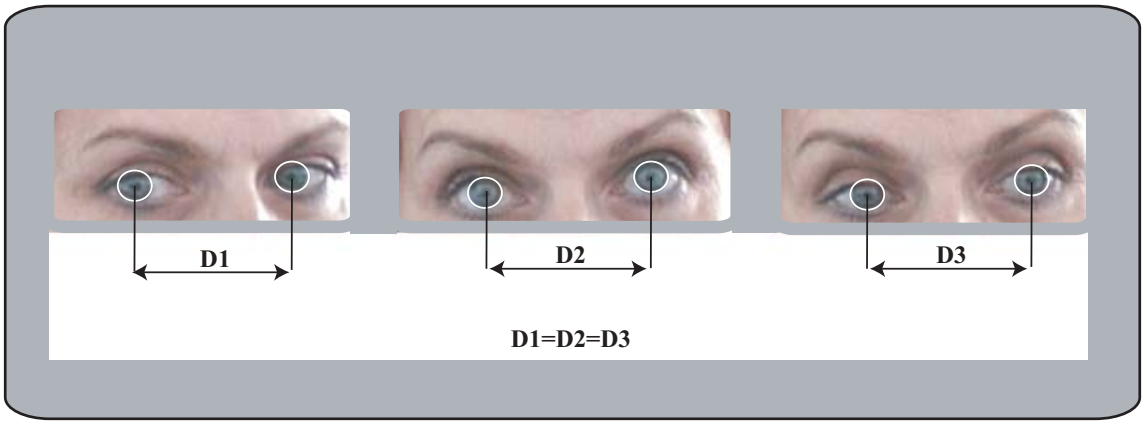

Figure 1.8: Distance between irises centers on different video sequences

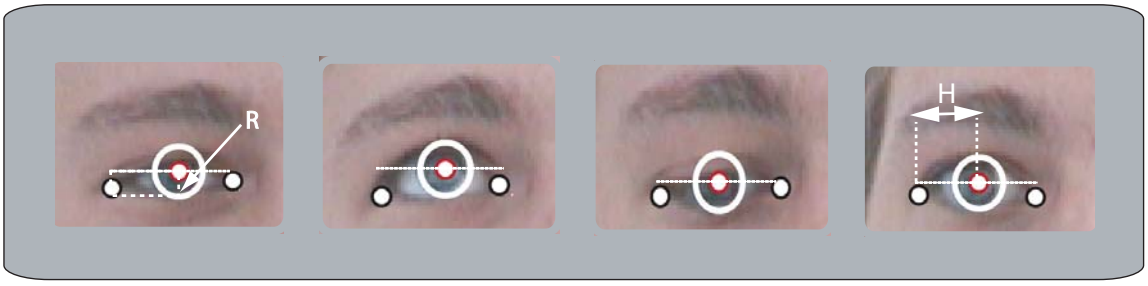

Figure 1.9: Positions of eyes corners relatively to iris center

- The horizontal distance H between the iris center and each corner is lower or equal to $3 * R$.

1.2.5 Summary

Figure 1.10 summarizes all the morphological constraints defined in this chapter and used in the segmentation process.

- (a): iris positions;
- (b): eyebrows bounding boxes positions and dimensions;
- (c): eyes bounding boxes positions and dimensions;
- (d): distance between irises centers;
- (e): eyes corners positions relatively to iris centers;

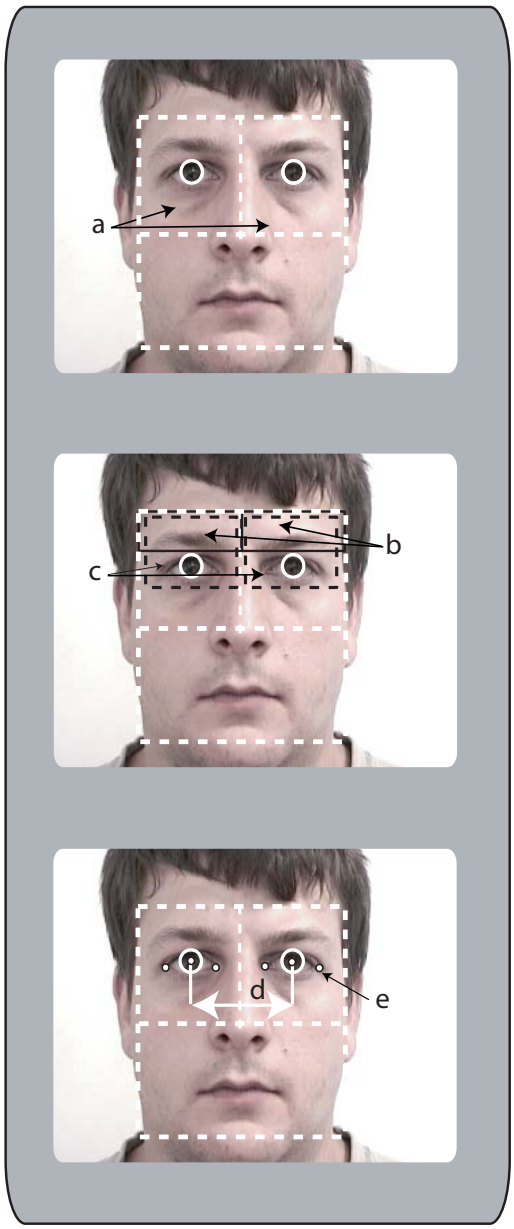

Figure 1.10: Morphological constraints.

Chapter 2

Iris, eyes, eyebrows and mouth segmentation

2.1 Introduction

The greatest difficulties encountered in detection of irises, eyes, eyebrows and mouth is caused by the luminance conditions, the complexity and high variability of the face morphology (color, form, presence of spectacles), facial expressions variations and head pose variations (zoom, pan and roll).

For the purpose of mouth segmentation, we use the algorithm developed in [Eve04] [Eve03a]. Here we only present algorithms for eyes and eyebrows segmentation.

During the last ten years, more and more works have tackled the problem of irises, eyes and eyebrows detection and/or segmentation. According to the extracted information and to some specific constraints, eyes and eyebrows segmentation methods can be classified into three main approaches: luminance and/or chrominance based approaches, active contours and deformable templates. The first approach extracts a coarse localization of these features based on luminance and/or chrominance information (valley images for example) ([Tse98], [Ko 99], [Den04], [Wan05], [Fas05], [Smi03]). The second approach is based on active contours or snakes which consist in deforming curves according to a cost function to be minimized [Coo98], [Rad95], [Par00], [Par01]. The third method is based on deformable templates which are matched to a person's eyes and eyebrows by minimization of a cost function [Yui92], [Kap98], [Tia00], [Mal01], [Vez03], [Mat00], [Kas01], [Dor04b].

2.1.1 Luminance and/or chrominance based approaches

The Luminance and/or chrominance based approaches can be divided into two main methods: training based methods and features based methods. Training based methods are not designed to achieve a fine segmentation of the facial features contours. So, as it is the purpose of our work, we do not describe these methods here.

Binarisation can be looked as an advantageous solution to highlight the facial features of the face while removing any useless information. But to obtain an accurate detection it requires a post-processing step to remove noisy results. In the case of eyes and eyebrows segmentation, it may be in their gradient properties, their morphological constraints, etc.

In [Kap98], the luminance information in a local defined area is used to characterize the eyebrows region. After the detection of the eyes centers, a search areas for the eyebrows

is defined above them. The eyebrows are detected assuming that they are darker than the surrounding skin. Then, a segmentation is carried out by binarization of the luminance image inside the search area. In Figure 2.1 eyebrows segmentation results are characterized by white pixels area. This method is only based on the binarization process inside an approximate area, then the results are very sensitive to the threshold value which is very difficult to define.

Figure 2.1: Results of eyes and eyebrows segmentation obtained with Kampmann method [Kap98].

There have been several observations about eyes and eyebrows gradient properties. D. Maio and D. Maltoni [Mai00] pointed out that eyes and eyebrows possess strong horizontal edges (for vertical faces). So eyes and eyebrows detection can be based on horizontal and vertical projection of the intensity gradient of the face.

In [Tse98] eyes and eyebrows are searched in the upper half of the face. The y-position of eyes is initially evaluated by the localization of the minima of luminance gradient of the vertical face profile. The set of minima whose distance in pixels lies in a predefined range, with respect to the width of the face, is chosen as being the eyebrows and the eyes. Then the horizontal eye gray level profile is evaluated. The maximum value of this profile is considered as the x-center of the face, and the minima from each side of the center whose distances from it are almost equal are considered as the x-locations of the eyes. Initial rectangular eye blocks are centered around the initially estimated eye positions and tracked by block matching in the remaining of the sequence. At each time the irises centers correspond to the centers of the blocks where a defined Matching Error (ME) is minimum. The head rotation is then computed according to the horizontal and vertical distances of the eye centers. Once the head rotated, eyebrows are searched with the same principle as the eyes detection. Their research area corresponds to the upper half of the face with a predefined range of distances before the known eye y-positions. Figure 2.2 shows an example of eyes and eyebrows segmentation.

The main disadvantage of this method is its dependence to the detection made in the first frame. Indeed this one is only based on the detection of the minima of the luminance gradient which sometimes leads to false detection results. This need the definition of additional constraints to eliminate them. However, these constraints are not always verified (for example x-center of the face detection). Then this method can not achieve robust result over different examples, head position and over different luminance conditions.

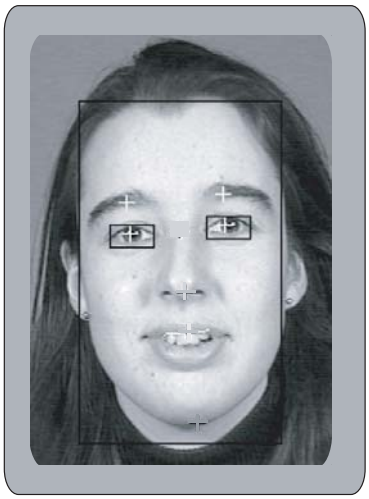

Figure 2.2: Segmentation results [Tse98].

Ko and all [Ko 99] added others properties for the eyes detection process. Their method is based on an adaptive thresholding followed by a selection based on morphological constraints. At the beginning, the image is binarized using a suitable threshold to enhance the eyes. The binary image is computed thanks to a heuristic thresholding method (P-Tile [Shi94]). A binary example of the used image is shown on Figure 2.3 top right. One label is associated to each connected block of the binary image. Morphological constraints are then used to select a set of candidate blocks. For each labeled block L two parameters are computed:

$$Size(L) = \sum_{x=x_1}^{x_2} \sum_{y=y_1}^{y_2} F(l(x,y)) \tag{2.1}$$

$$Ratio(L) = Hight_{bloc}/Width_{bloc} \tag{2.2}$$

where (x_1, y_1) are the coordinates of the higher left corner, (x_2, y_2), the coordinates of the lower right corner, $l(x, y)$ is the label of the pixel (x, y) and $F(i) = 1$ if $l(x, y) = L$; $Size(L)$ corresponds to a criterion on the eyes size in pixels, while $Ratio(L)$ characterizes the eyes form. Based on these two parameters, the selected blocks candidates are those which check the two following conditions:

$$Min \leq Size(L) \leq Max \quad \text{and} \quad Ratio(L) \leq 1 \tag{2.3}$$

Min and Max are learned constants.

Once the false detections reduced, the blocks candidates are gathered computing the similarities (size and form) between pairwise of block candidates. The two most similar blocks are selected and are assumed to be the eyes (see Figure 2.3.c).

The limitation of the method is its dependency to the two parameters Min and Max. The two parameters are very dependent to the luminance conditions, the distance of the subject to the camera and the binarisation process during the training process. Then the selection process leads sometimes to false detection results.

Instead of using a predefined threshold to improve the selection process, [Den04] combines the intensity and the luminance gradient. In each detected face, the horizontal projection of

Figure 2.3: Eyes localization. Top: left, frame of face; right, binarized image. Bottom: localized eyes [Ko 99].

the vertical contours and the horizontal projection of the intensity are computed. The top two peaks of the horizontal projection are detected and the one with the lower intensity is defined as the y coordinate of the eyes (see Figure 2.4.a). Based on this coordinate the eye region is cropped from the face region and the vertical projection of the horizontal contours is computed. The detected top two peaks are considered as being the x coordinates of the eyes (see Figure 2.4.b). Once the locations of the eyes are given, two iterations of Otsu's thresholding [Ots78] are then applied to obtain a binary image. Finally a window growing is used to extract the eyes bounding boxes. The parameters required in window growing are the centroid of the window, the initial size of the window and the size of growing step. No information is given about the definition of these parameters nor about the growing stop criterion. An example of cropped eye region is shown in Figure 2.4.c. The method is sensitive to glasses. Moreover, it presents a limitation to men's mustache and women's long hair since both mustache and hair can generate vertical edges and possess low intensity values which leads to false detection results.

In [Smi03] eye detection is based on eyes color information. First, the skin color predicate is built using [Kje96], which segments the skin from non-skin regions. Since eyes are not skin pixels, they always show up as holes. The system finds the two holes that are above the lip region (segmented in a previous step) and that satisfy a defined size criteria for eyes according to the face. Camera is at a fixed distance from the face so as to have a relative size of eyes to be between 1% and 2% of the area of the image. Figure 2.5 shows an input image (left), the output of the color predicate program (middle) and the output of the automatic eye initialization (right). In the remaining of the sequence eye tracking is done in a multi-strategy way. First the system uses intensity information of the eye region to find eye pupils. They correspond to the center of mass of the eye region pixels. Then a small window is searched. An affine motion model is applied to make the tracking of this small window around the eyes from the current frame to the next frame.

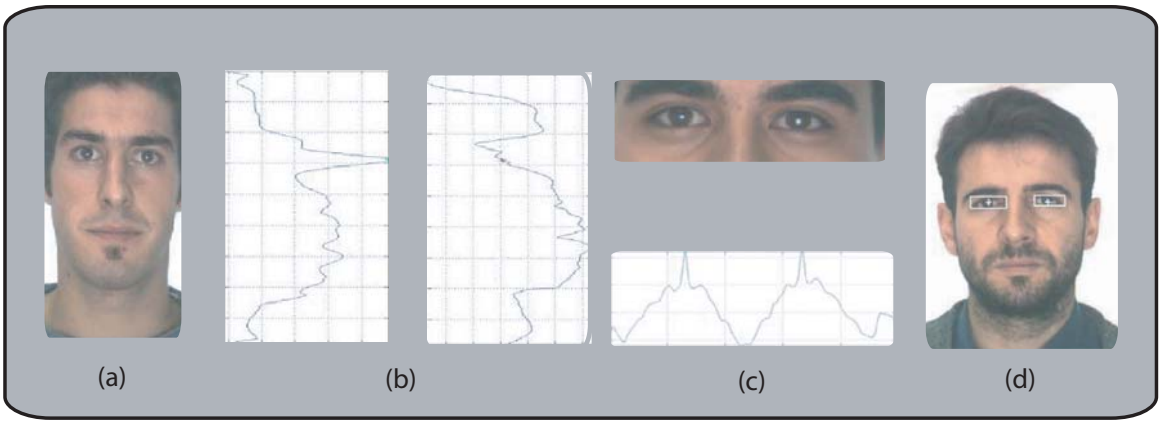

Figure 2.4: (a): input image; (b): horizontal projection of horizontal transitions (vertical edges) and of intensity, (c): vertical projection of vertical transitions subtracted by intensity, (d): extracted eyes in the image [Den04].

Figure 2.5: Left: input image; middle: selected skin region; right: results from automatic eye initialization [Smi03].

The previous techniques do not obtain a regularized segmentation of the feature contours. To refine the segmentation results, [Vez03] introduces an eye contour model based on a circle for the iris, a cubic curve for the upper eyelid and on a quadratic curve for the lower eyelid. Their method is based on a color image containing a single human eye and performed in three steps: approximate detection of the eye center, extraction of iris shape and extraction of eyelid curves. The approximate eye center is found based on the red channel. According to empirical threshold and a set of luminance distribution conditions, a set of circular window are selected. Their mean coordinates determine the approximate eye center. The iris center coordinates are refined during the detection of the exact iris radius found by searching for a circle which lies on the border between dark pixels of iris and bright pixel of the eye white. The upper eyelid is found by looking for luminance valley points corresponding to points of significant local minima of the horizontal luminance profiles (see Figure 2.6.a). Outliers points, corresponding to false detection, are rejected using a Hough transform (see Figure 2.6.b). The line with maximum number of points lying closer than a predefined distance is chosen. Points that lie

too far from the main line are removed from the set of boundary points. Then leftmost and rightmost points are chosen to be the eye corners (see Figure 2.6.c). Finally, a cubic curve is fitted to the correct eyelid points. The lower eyelid is estimated by fitting a quadratic curve to the eye corners and the lower point of the detected iris circle. This method is very sensitive to the luminance conditions and the need of a precise knowledge of the eyes regions. The scale and the eye bounding box are considered to be known. The definition of the thresholds and of the parameters used in the segmentation process depend on the accuracy of this area.

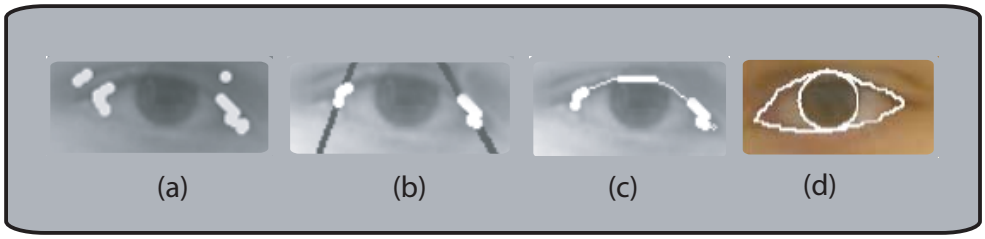

Figure 2.6: (a) Initial eye border points set (luminance valley points); (b) set with marked principle lines outliers removed; (c) cubic polynomial curve fitted to the final border points set; (d) segmentation result [Vez03].

[Wan05] argues that working on images with the full face makes the eye details small and affects the accuracy. The authors address this problem by zooming in on a single eye. The iris contour is modeled by an ellipse and to extract it the first step consists in detecting the upper and lower eyelids. The steps of the iris detection process are shown in Figure 2.7. First, the original image with a single eye (see Figure 2.7.a) is adaptively segmented based on the histogram to yield the binarised image (see Figure 2.7.b). Then morphological operators allow to obtain the results of the Figure 2.7.c. A vertical (Canny) edge operator result is shown in Figure 2.7.d. The two longest edges are used for fitting the iris contour (see Figure 2.7.e) and their superimposition is shown in Figure 2.7.f. Hence the two outer boundaries of the iris that are not occluded by the eyelids are used for ellipse fitting. The resulting ellipse can be seen in Figure 2.7.g and its superimposition onto the eye in Figure 2.7.h. The proposed method needs either a camera dedicated for eye detection or a precise eye bounding box detection leading to area such as Figure 2.7.a. Moreover, dedicated to the gaze estimation, it only detect the iris.

[Kas01] proposes eyes region detection and an iris segmentation method based on hybrid features. The face region is obtained by the color difference from the standard skin color UV. Secondly, eyes regions are extracted from the face region by hybrid template matching based on edge (the horizontal, the vertical, the top right, and the top left edge frames) and the color (UV space) distance features. Facial parts can be extracted against individual variation of faces and head movements. Finally, the iris regions are detected by using saturation and brightness from the extracted eye regions (see Figure 2.8.a). First, the extracted eye region is divided into the eye and skin regions by a saturation histogram (see Figure 2.8.b). Then, the eye region is divided into the iris and the white of the eye from the brightness histogram. Next, the edge of the iris is detected from the gray-scale image in the iris region by Prewitt's operator (see Figure 2.8.c). Finally, the iris boundaries are detected using the Hough Transform.

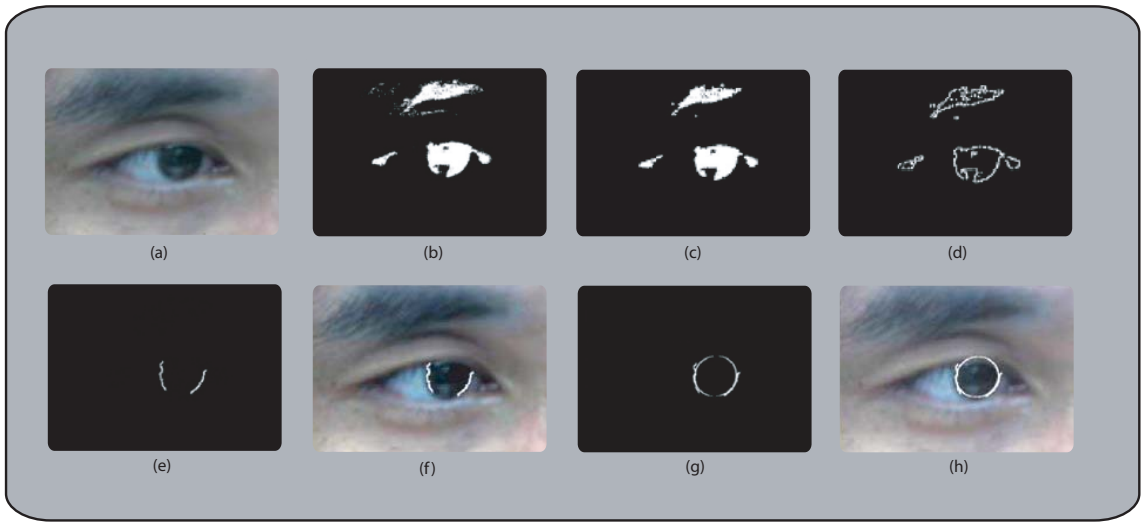

Figure 2.7: Iris detection: (a) original image; (b) thresholding results; (c) morphological *open* operation; (d) vertical edges; (e) two longest edges by region-following; (f) overlay edges onto the original image; (g) edge result and least-squares fitted ellipse and (h) overlay edge result and least-squares fitted ellipse onto the original image [Wan05].

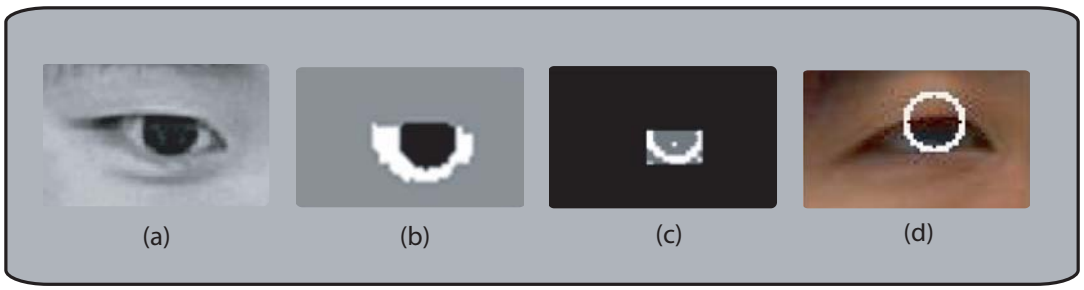

Figure 2.8: (a): Extracted eye region; (b): segmented skin and eye; (c): detected edges of the iris in the iris region; (d): iris detection [Kas01].

Luminance based algorithms need the tuning of thresholds and are sensitive to lighting variations. The gray level information is helpful in detecting several eye candidates but using it alone is not sufficient to filter out other facial features such as eyebrows and mouth, which also appear as dark patches. Another disadvantage is the poor extracted information (for example region and not contour).

2.1.2 Active contours based approaches

Introduced by [Kas88], Snakes (active contours) have been accepted as a standard technique to extract flexible contours. A snake is defined as an energy minimizing spline. The snake's energy depends on its shape and location within the image. A snake is a parametric curve c ($c(s) = (x(s), y(s)$ with s being the curvilinear abscissa) which can be iteratively deformed so as to fit the contours of the segmented object. The deformation is oriented by the minimization

of a cost function composed of two terms: an internal energy function and an external energy function which establishes interaction constraints to maintain the consistency between the template geometry and the relevant image features.

For eyes and eyebrows segmentation Radeva and all in [Rad95] propose to use the rubber snakes [Rad93] models and assume to have a more accurate and realistic representation of the detected feature shape. In order to make the detection more robust they add a constraint of symmetry on the face. Based on the horizontal and vertical luminance projection, eyes and eyebrows positions are approximatively detected. Then eyes and eyebrows models are also approximatively initialized. However, very few details are given on the kind of models used and their initialization. The active contours process needs specific constraints to guide their evolution which can be introduced modeling hypothesis on the searched shapes. In order to increase the global control of the snake shape deformation for the eyelids segmentation they add a new term to the internal energy function of the used rubber snake. It allows to take into account the symmetry between both eyelashes. An example of the Radeva *et al* results is shown in Figure 2.9.

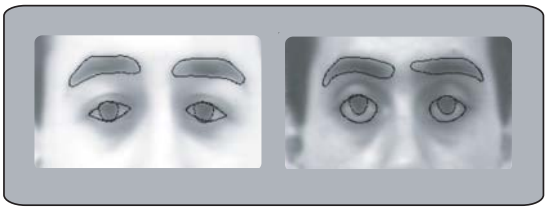

Figure 2.9: Examples of segmented eyes and eyebrows using rubber snakes [Rad95].

Based on the same idea (using luminance for eyes localization) [Par00] and [Par01] manage to segment eyelids and eyebrows boundaries. In [Par00] the minimal paths algorithm is used to find the eyelids in the first frame where the eyes are presumably located and a snake tracking method is used for the remaining frames of the sequence. To restrict this area, morphological operators are used to define the dark contrasted components (which usually include the pupils and eyelids see Figure 2.10.b) and the white contrasted components (the white of the eyes see Figure 2.10.c) which are used as search areas for the eyes segmentation. A first approximation of eyes corner points is then extracted using deformable line templates [Ami90]. Then the minimal path algorithm described in [Par00] is applied for each pairwise of candidate corner points. The pairwise of corner points leading to the minimal path is selected as the correct one and this path as the eyelids contours of the first frame (see Figure 2.10.d).

For the next frames, first a closed snake is built by selecting a small percentage of points from the detected eyelids contours [Par00] in the first frame. Based on the segmentation results of the eyebrows defined in [Par99] closed or open snakes are also defined by selecting a small percentage of points. Then, motion estimation is introduced in the energy minimization procedure to track it [Par01]. To do this, a small and rectangular region around each pixel is selected as a basis for the motion estimation. A motion error called the Motion Compensation Error (MCE) is computed for all the possible displacements of the block in a given range and added to the external energy. The main limitation of this approach is the initialization step. It must be as close as possible to the required contours which is not an easy task. Moreover, the snakes are *free form models* and do not integrate any information about the required forms. This implies the addition of a specific constraint to guide their evolution.

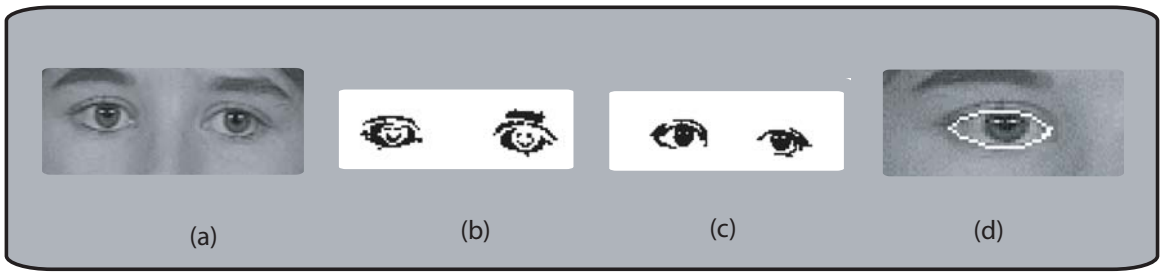

Figure 2.10: (a): Original image; (b): dark contrasted components; (c): clear contrasted components; (d): detection result [Par00].

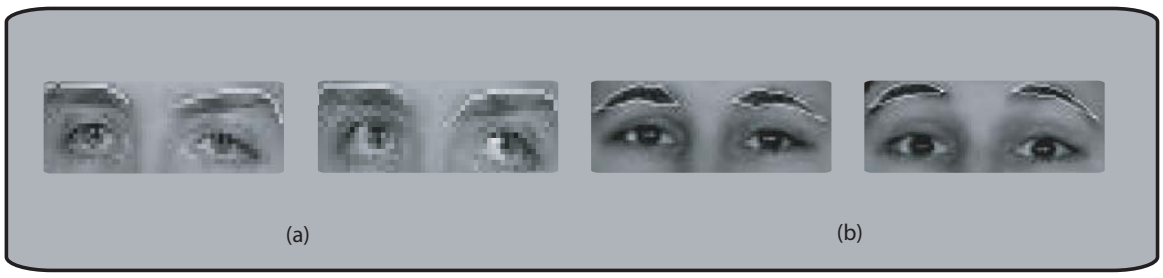

Figure 2.11: (a): Tracking results with open snakes; (b): tracking results with closed snakes [Par01].

2.1.3 Deformable templates based approaches

Introduced by Yuille [Yui92] the deformable templates are specified by a set of parameters which enable a priori knowledge about the expected shape of the features to guide their deformation process. The eyes model chosen by Yuille is composed of a circle for the iris and two parabolas for the two eyelids (see Figure 2.12). The deformable templates have some similarities with the snakes, but, in addition they have the advantage to explicitly evolve under the constraints of a specific model, contrary to the snake evolution which is a blind process which gives no guarantee on the validity of the final shape. In Yuille's method the eyes components are linked together by three forces corresponding to the internal energy: first, forces which encourage the center of the iris and the eye bounding contour to be close together; second, forces which make the eye width roughly four times the radius of the iris; and third, forces which encourage the centers of the eyes whites to be roughly midway from the center of the eye to the boundary.

A great number of studies have focused on the use of deformable templates. In order to make the segmentation independent on the person morphology and to make the minimisation step faster, Kampmann *et al* [Kap98] propose a segmentation method based on a simplified cost function. They use Yuille's eyes model (Figure 2.12). Firstly the pupil and the corner points of the eyes are estimated by [Zha97] and [Zha96]. Based on this position, a search window is defined for eyelids opening's height estimation (maximal distance between the eyelids). The simplified cost function does not impose any constraints on the eye features and is based on spatial luminance gradients along eyelids and iris edges as well as means and variances

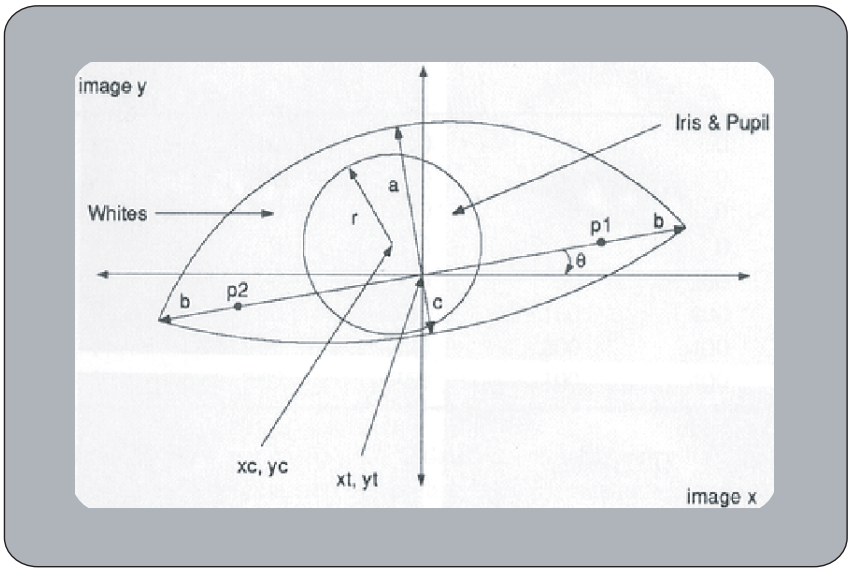

Figure 2.12: The eye model defined by Yuille: two parabolas for the eyelids and a circle for the iris [Yui92].

of the luminance in eye white and iris areas. At first, eyelid opening's height candidates are determined by minimizing the cost function considering only the luminance gradients. Then a verification step allows to estimate the most accurate candidate. No explanation or information is given about the verification criteria. Figure 2.1 shows an example of Kampman's eyes segmentation.

Malciu *et al* [Mal01] also propose a method to improve the accuracy of the deformable template segmentation. The internal energy function for the eye template is designed to in-

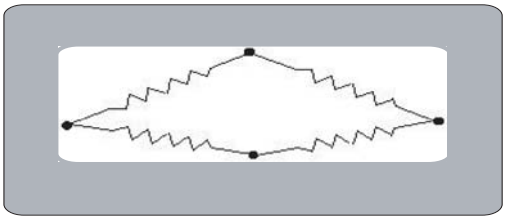

Figure 2.13: Eye template: elastic internal energy model [Mal01].

corporate rigidity and local symmetry constraints. The template is a system made of ideal springs connecting neighboring nodal points along each one of their curvilinear components (see Figure 2.13). The constraint is that the middle point along each eye boundary remains in a central position. The external energy function is meant to maintain the consistency between the template geometry and relevant image features. The optimal template deformation (parametrically defined as a second order polynomial expression) is estimated by minimizing the total energy function using the simplex method [Prê92].

The fastest algorithms are those which do not use iterative optimization techniques to minimize the cost function. This approach has been adapted by Botino *et al* [Bot02], Tian

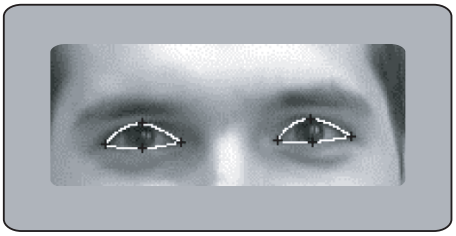

Figure 2.14: Results of Malciu segmentation method [Mal01].

et al [Tia00] and Dornaika *et al* [Dor04a]. In [Bot02] and [Tia00] the eyebrows model is composed of three characteristic points with two connecting segments. For the eyes a multi-state template is defined. The open eye model corresponds to two parabolas for the eyelids and a circle for the iris (nine parameters). The closed eye is modeled by one line between the eye corners (four parameters) ([Bot02], [Tia00])(see Figure 2.15). The detection is based on the manual initialization in the first frame of the eyes and eyebrows characteristic points. These points are then tracked in the remaining frames of the sequence. To do this, Lucas-Kanade tracking algorithm [Luc81] is used in [Tia00] whereas in [Bot02] a correction step based on defined template models is associated to the tracking process.

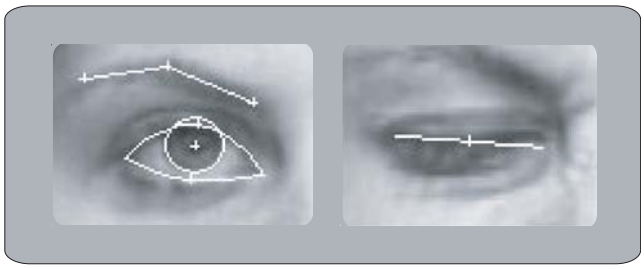

Figure 2.15: Eyes models. Left: open eye; right: closed eye [Tia00].

Figure 2.16: Tracking results by Tian and al [Tia00].

These methods, based on the use of characteristic points, are probably the fastest ones but they present two main drawbacks. Firstly, the use of only the position of some points (local information) to compute the model is not a robust method to noise contrary to the

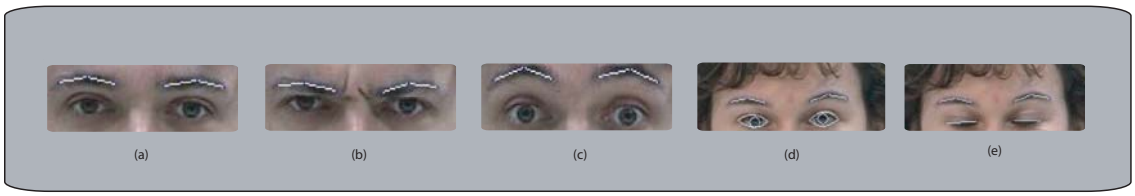

Figure 2.17: Eyebrows tracking results by Botino [Bot02].

computation of the integrals along the curves (global information). Secondly these methods are completely based on the tracking points algorithm accuracy. Notably their precision strongly decrease when the points are located in very deformable zones or when the contours are very broadly defined.

2.1.4 Conclusion

It is difficult to make an objective comparison between all these methods. The performances of each of them depend on the type of applications and the acquisition constraints of the used images. Nevertheless, we can summarize their advantages and drawbacks.

The luminance information based approaches only rely on the use of pixels information. They suppose that eyes and eyebrows have sufficiently salient luminance characteristics to differentiate them to the other facial features. They assume that eyes and eyebrows are the darkest zones in the face and that eyes contain strong horizontal edges. Based on this information, intensity thresholding can be applied to extract them. However defining thresholds is a highly sensitive approach and strongly depends on the acquisition conditions and especially on the lighting conditions. To overcome these problems other information is added to filter out noisy results combining luminance gradient and intensity information and introducing morphological constraints. Finally these methods allow a fast localization of the features but they do not allow to obtain an accurate and precise detection of the contours.

The second approach is based on active contours (snakes). They also make use of the luminance information in the image (gradient, peak and valley), they introduce regularity constraints like symmetry and elasticity of the segmented contours. The main advantage of the snakes is their great ability to handle deformations and adaptations on a great number of objects contours. However, their main limitation is the initialization step. Indeed, the forces which deform them only depend on their near neighborhood. Then, if the contour is initialized too far from this final contour, it has few chances to join it and can deviate from the real targeted contour. It is possible to increase the elasticity and curvature constraints of the detected shape but this leads to an increase of the computation time. The deformable templates approaches add shape constraints to the snakes. Using this approach, the shape of the searched features is always valid. Moreover as the convergence process is applied on the model globally, the influence of noise or of local disturbances is less important. However, shape constraints lead sometimes to less accurate segmentation results and do not allow to adapt to a great number of deformations like the ones occurring on eyes and eyebrows in the case of facial expressions. Initialization not too far from the final position is also necessary for good detection results.

In the next section we present a method for eyes and eyebrows segmentation which uses a mixture of information shared by the three described techniques. So as to be applicable in facial expressions recognition, contours must be extracted with maximum precision. First of all, we will adopt the same idea as the luminance based techniques using the property that iris, eyes and eyebrows contours are characterized by a maximum luminance gradient. However in order to overcome the limitations of these methods to the luminance conditions and noise information a preprocessing step based on human retina modeling is applied for illumination variation removing. A set of morphological constraints of the face are used to define the research area of eyes and eyebrows. The robustness of our method will be ensured by the use of a deformable model for eyes and eyebrows. A set of key points is detected in the neighborhood of the required features to initialize the model. Then the maximization of luminance gradient is used to fit them. Finally we also introduce the lips segmentation method [Eve03a] in order to obtain a complete facial feature extraction system. The segmentation algorithm is based on the use of color information and is based on an analytical mouth deformable model, flexible enough to reproduce a wide range of shapes.

2.2 Iris segmentation algorithm

Face extraction is beyond the scope of our work. Several algorithms have been proposed for the automatic extraction of the face in images (see the survey articles [Yan99], [Hje01]). The proposed method imposes in the first image of the sequence, a manual extraction of the face bounding box (see Figure 2.22). Then the face is automatically tracked by block matching in the remaining frames of the sequence.

2.2.1 Retinal preprocessing

One of the major problem in image processing is the development of methods robust towards illumination variations. To avoid these problems, a preprocessing stage based on the model of human retina [Bea94] [Hér96] [Tor99] is applied. The retina presents the first set of filtering processes which transforms and prepares the visual input information to be analyzed by higher cortical areas. Here, we propose to apply this model to enhance the contours (the luminance gradients) and at the same time, to realize a local correction of the illumination variations. The retinal processing can be described as a multi-stage combination (see Figure 2.18).

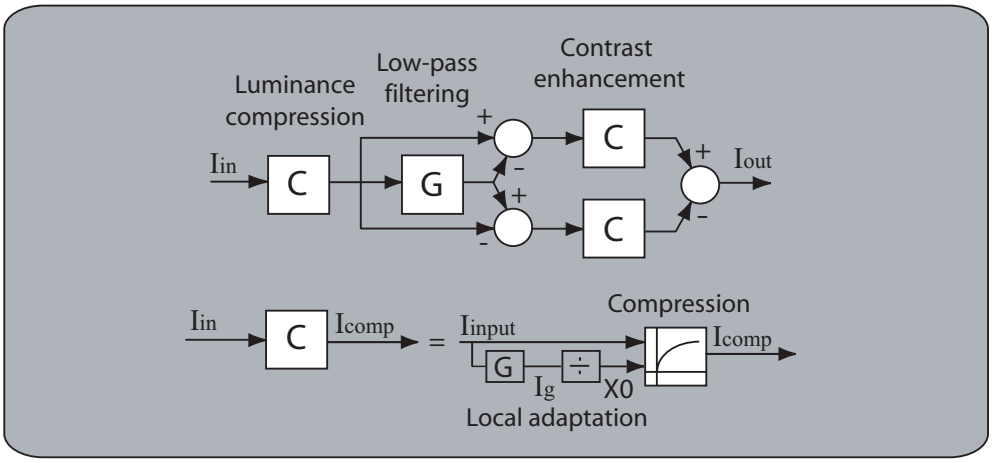

Figure 2.18: The different stages of the retinal preprocessing.

Figure 2.19 describes the different stages of the retinal preprocessing:

- (a): input frame I_{in} (see Figure 2.19.a).
- (b): a compression is applied on each pixel of the original image I_{in} leading to a suppression of large luminance variations. This compression varies according to the local intensity so as to obtain a weak compression in bright regions and a strong compression in dark regions. This local adaptation is done throught a parameter X_0 which depends on the mean intensity inside a small region around each considered pixel (see Figure 2.19.b).

  For each pixel p of I_{in} the compression is defined by (see Figure 2.20):

$$I_{comp}(p) = \frac{(255 + X_0(p))I_{in}(p)}{X_0(p) + I_{in}(p)} \tag{2.4}$$

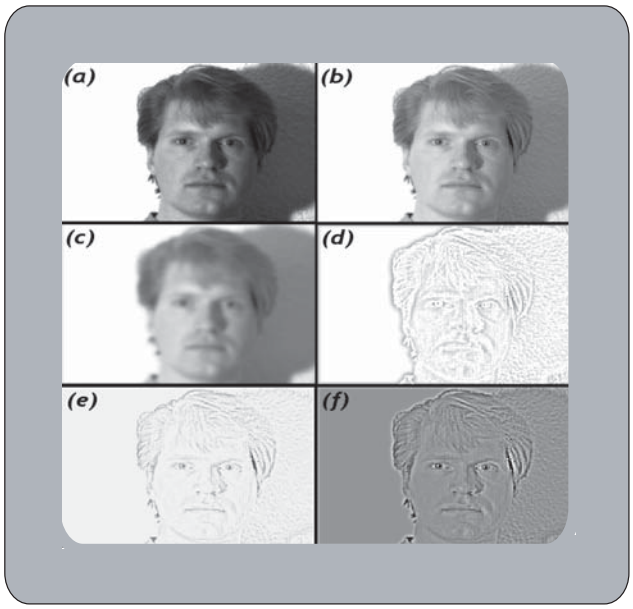

Figure 2.19: The chronological outputs of the retinal preprocessing.

with X_0 computed according to I_g, the filtered version of I_{in} (using a gaussian filter

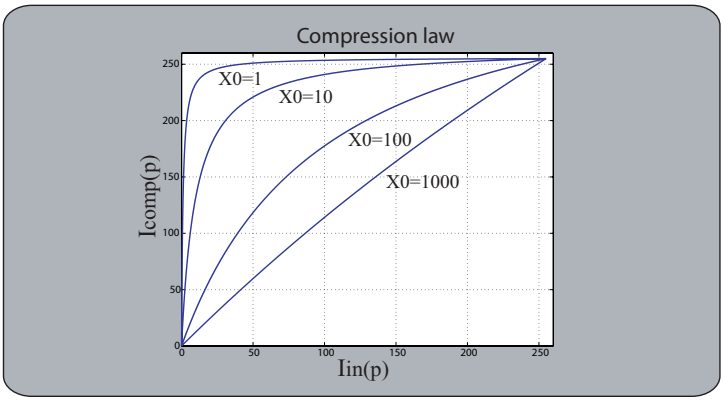

Figure 2.20: Compression sheme (X0 being the local mean of luminance).

with a size of 15x15 pixels and a standard deviation of 2). For each pixel p, X_0 is obtained as:

$$X_0(p) = 0.1 + \frac{410 I_g(p)}{105.5 + I_g(p)} \tag{2.5}$$

- (c): a local averaging is done using a gaussian filter G on the output of stage (a) (see Figure 2.19.c).

- (d) (e): two opposite differences of the outputs of stages (b) and (c) are then realized and a second local compression is applied on both outputs leading to an enhancement of the local contrasts (see Figure 2.19.d and Figure 2.19.e).
- (f): a final summation of outputs of (d) and (e) is done leading to the final output (see Figure 2.19.f).

In the final output of the retinal preprocessing (see Figure 2.19.f) the image contrast has been enhanced and local luminance variations has been removed.

2.2.2 Luminance gradient for iris segmentation

Iris contour is the frontier between the dark area of iris and the eye white. We choose to use a circle as iris model which proves to be sufficient for the purpose of facial expressions classification. This contour is supposed to be a circle made of points of maximum luminance gradient. Since the eyes could be slightly closed, the upper part of the iris could be occluded. So for each iris, we are looking for the lower part of the iris circle.

The luminance gradient on face bounding box is computed by using the Sobel operator which gives the image gradient $\overrightarrow{\nabla I_t}$. One representation of the gradient vector (norm and orientation) at each point of the frame is displayed in Figure 2.22.c.

The luminance gradient is computed and the morphological constraints described in section 1.2.2 are used to select a bounding box around each eye (see Figure 2.22.c). In each eye bounding box, each iris semi-circle maximizes the normalized flow of luminance gradient, noted $NFLG_t$:

$$NFLG_t = \frac{1}{length(C_{sc})} \sum_{p \in C_{sc}} \vec{\nabla} I_t(p).\overrightarrow{n}(p) \tag{2.6}$$

where $I_t(p)$ is the luminance at point p and at time t, $\vec{n}(p)$ is the normal to the boundary at point p and C_{sc} is the boundary of the lower semi-circle.

The $NFLG_t$ is normalized by the length of C_{sc}. In order to select the semi-circle which

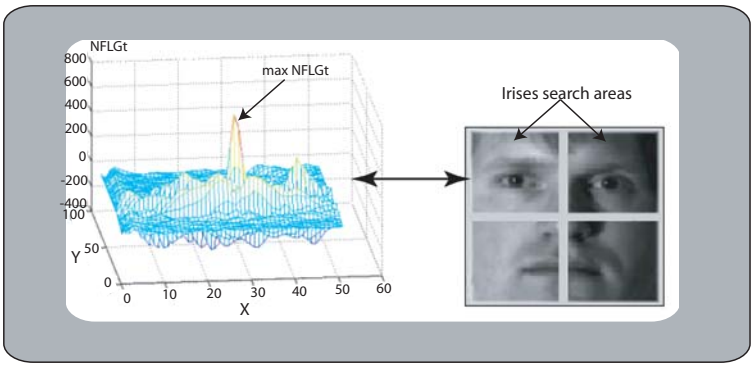

Figure 2.21: Evolution of the $NFLG_t$ during the scanning of the iris search area corresponding to the upper left square of the rectangle surrounding the face.

maximizes the $NFLG_t$, several candidates semi-circles are tested during a scanning process

of the search area of each iris. Figure 2.21 shows that the right position of the semi-circle exhibits a sharp maximum of $NFLG_t$ in the iris search area. The detected semi-circle are then completed by symmetry (see Figure 2.22.d).

Iris radius is supposed to be known and only the center position of the searched semi-circle is scanning the eye bounding box. It could be possible to solve this manual estimation by automatically testing several radius values since this one is correlated to the face dimensions.

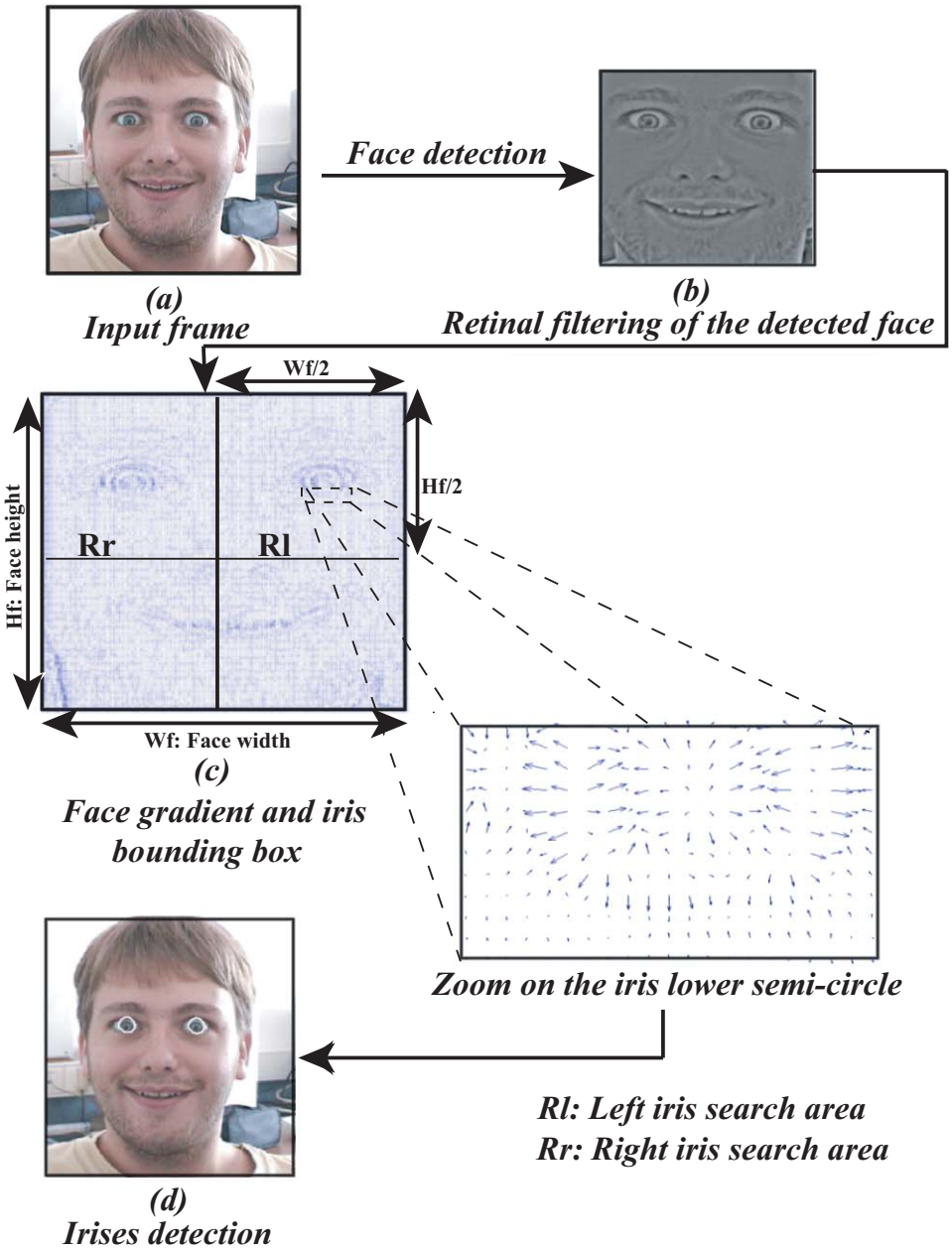

Figure 2.22: Iris segmentation process: (a) input frame; (b) retinal preprocessing; (c) face gradient and zoom on the iris semi circle area; (d) final iris detection.

In order to be robust to extreme case of lighting, we also use the constraints of the section 1.2.3: in the time course the distance between the irises centers is constant as long as the two irises are completely visible (maximum pan used as showed in Figure 2.30). However this distance decreases until ($3/2 * R$ (R being iris radius)) from a pan equal to 40^o which corresponds to the limit of our segmentation algorithm (see section 1.2.3). Thus, at each new detection, this distance is computed to validate the current detection. The used reference is the distance between the two irises in the first frame. In the case of error the last position is used. Figure 2.23 shows the detection of the left iris after a correction process.

Figure 2.23: Correction of the iris false detection; left: false detection of the left iris; right: detection after correction.

2.2.2.1 Eyes states

A blink corresponds to the transition of the state of each eye from *open eye* to *closed eye*. The *open* or *closed* state of each eye is related to the presence or the absence of an iris. The automatic detection of the eye state is based on the analysis of the $NFLG_t$ (Normalized Flow of Luminance Gradient). When the eyes start closing, the iris semi-circle is less and less visible so that $NFLG_t$ is decreasing. Sometimes when the eyes are closed the $NFLG_t$ value along the selected semi-circle when the eyes are closed is perturbated by the presence of the lashes: the semi-circle coincides with the lashes which are made of points of maximum gradient of luminance (frontier between the skin and the lashes). As a result, the $NFLG_t$ along the selected semi-circle corresponds to an open eye. For this reason, we add the normalized mean value of luminance of the semi-circle surface (NQL_t). Once the eye start closing this value increases. It is defined by the relation:

$$NQL_t = \frac{\sum_{p \in S_{sc}} I_t(p)/nbr}{sup_{p \in S_{sc}} I_t(p)} \tag{2.7}$$

where $I_t(p)$ is the luminance at pixel p at time t, S_{sc} is the surface of the candidate semi-circle and nbr the number of its points (pixels).

Let $NFLG_m$ and NQL_m be the mean values of the $NFLG_t$ and the NQL_t for the open eyes, computed on a temporal window with a width Δ_t and situated between $t - \Delta_t$ and t. Δ_t corresponds to the time needed for one or two blinks at normal blinking frequency and is equal to 150 frames at 25 frames/s. This value corresponds to the values found in the medical literature [Bli02](12 to 20 blinks per minute). The evaluation of $NFLG_m$ and NQL_m

at different time t allows the system to re-adapt itself to varying acquisition conditions (like change in illumination conditions). At time t, the maximum of $NFLG_m$ and the minimum of NQL_m overall the already estimated values inside the defined temporal window are computed. Indeed $NFLG_t$ depends on the degree of opening of the eye (Figure 2.25 right) and taking the maximum value ensures to consider the $NFLG_t$ corresponding to the highest opening of the eye. On the contrary, as the iris is a dark area NQL_t decreases with the opening of the eye (Figure 2.25 left) and taking the minimum value ensures to consider the NQL_t of the highest opening of the eye. Then at each frame t, eyes are detected as open if the following relations are satisfied:

$$\begin{aligned} &(NFLG_t \geq max(NFLG_m) * c_{NFLG}) \\ &and(NQL_t \leq min(NQL_m) * c_{NQL})) \end{aligned} \tag{2.8}$$

Once the eyes are closed, the value of $NFLG_t$ is inferior to the maximum of $NFLG_m$ multiplied by a coefficient c_{NFLG} (first condition for closed eyes). If the eyes are closed, NQL_t is higher or equal to the minimum of NQL_m multiplied by a coefficient c_{NQL} (second condition for closed eyes) because the surface of the selected semi-circle corresponds to a clear area (area of eyelids) in the case of closed eye, instead of a dark area (area of iris) in the case of open eye. The coefficients c_{NFLG} and c_{NQL} are taken, so that an eye is considered open if more than $\frac{1}{3}$ of the semi circle is visible (Figure 2.24 and Figure 2.26) and is considered closed otherwise. These coefficients are computed as the mean of their values on ten subjects and allow to obtain the ratio between the maximum and minimum value of $NFLG_t$ and of NQL_t from the open to closed eye state. Figure 2.25 shows the temporal evolution of $NFLG_t$ and

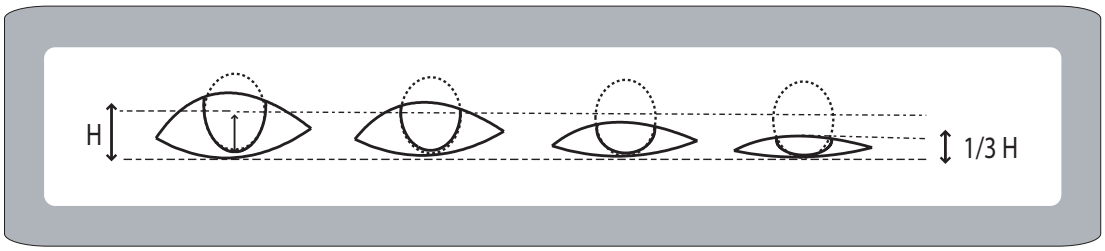

Figure 2.24: Evolution of iris opening in the case of blink sequence.

NQL_t and the results after thresholding. On these curves, there are two blinks: the first one is very quick and the second one occurs during several frames (it might correspond to a short sleeping). Figure 2.26 shows an example of iris segmentation during a blink.

2.3 Iris segmentation results

2.3.1 Robustness and precision of the iris segmentation process

Figure 2.27 shows different iris segmentation results: the first and second rows show the robustness of our algorithm to different facial expressions and ethnicity, the third row, to spectacles and bad luminance conditions. The analysis of the iris segmentation on the HCE (all the frames), HF (all the frames), YF (40 selected frames), CKE (144 selected frames) and FF (60 selected frames) databases gives satisfactory visual results on nearly all the frames.

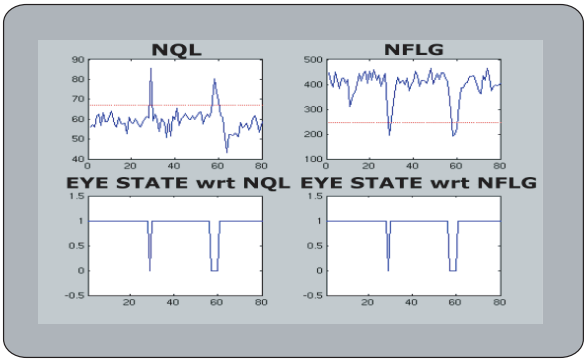

Figure 2.25: Top left: temporal evolution of NQL_t and threshold (dashed line); top right, temporal evolution of $NFLG_t$ and threshold (dashed line); bottom left, eye state after NQL_t thresholding; bottom right, eye state after $NFLG_t$ thresholding (0 stands for closed eye and 1 stands for open eye).

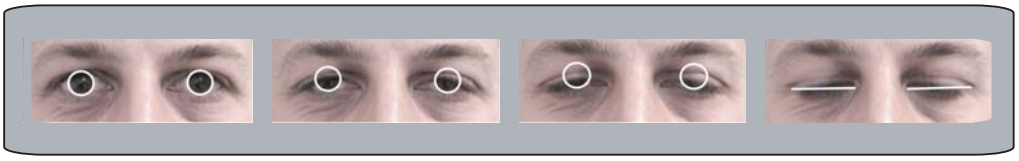

Figure 2.26: Sequence of iris segmentation results during a blink.

The detection of the iris can be performed also in a video sequence with head motion, with or without spectacles (sight glasses) in conditions where the iris is always visible in the face bounding box.

In order to have a quantitative evaluation of the iris detection a human validation of the automatic detection process has been done. To do this a *ground truth* is required. It consists in having the results of a manual iris segmentation by an expert. Moreover, the number of frames included in the reference set must be sufficiently important so that the measurement precision is significant. For the iris, it consists in measuring the precision of its center position. However this task appears to be difficult when performed by human observers because the iris is a dark area and lots of refinements have to be done iteratively. Hence we have used another kind of precision measurement based on a visual criteria.

The segmentation results obtained on frames from a test set have been manually validated as true or false detection. We have chosen a tolerance threshold of 2 pixels for the iris position. The segmentation result is selected as true if the detected contour is shifted with +/- 2 pixels around the true contour and false otherwise. Figure 2.28 presents an example of the iris contour and four examples of the possible shifted cases. Figure 2.28.a corresponds to the correct iris segmentation result. Figure 2.28.b, 2.28.c, 2.28.d and 2.28.e corresponds to 2 pixels shift of the iris contour towards respectively the top, the bottom, the left and the right of the true iris. The same validation process is used whatever the shift direction of the segmented iris contour.

The irises of 1830 frames coming from 8 different subjects from HCE database and 3 sequences of HF database have been used. They have been automatically segmented using the

Figure 2.27: Results of iris segmentation on the YF database (first row frames 1 and 2, second row frames 2 and 4), the CKE database (first row frames 3 and 4), the HCE database (second row frames 1 and 3) and the FF database (third row).

proposed algorithm. The HCE frames are selected such as for each one of the 8 subjects the 60 first frames of each one of their three expressive sequences are selected. Thus $60*3*8 = 1200$ frames from the HCE database are used in order to test the robustness of the segmentation process to facial expressions. The HF frames are selected such as the 130 first frames of each sequence corresponding to the 3 subjects. Thus $130 * 3 = 390$ frames from the HF database are used to test the robustness of the segmentation process to head and iris motion. Then a total of 1590 frames has been used to measure the iris segmentation precision with the iris radius varying from 6 pixels to 12 pixels.

A +/- 2 pixels shift corresponds to an error of 1 pixel on the iris radius. To evaluate the accuracy of the iris segmentation we have adopted a method maintaining the imprecision error constant in pixels. Our choice is justified by the fact that while the iris radius increases, the more the resolution of the image is and then the more the precision of the segmentation should be. As in our tests the iris diameter varied from 12 to 24 pixels then the error varied

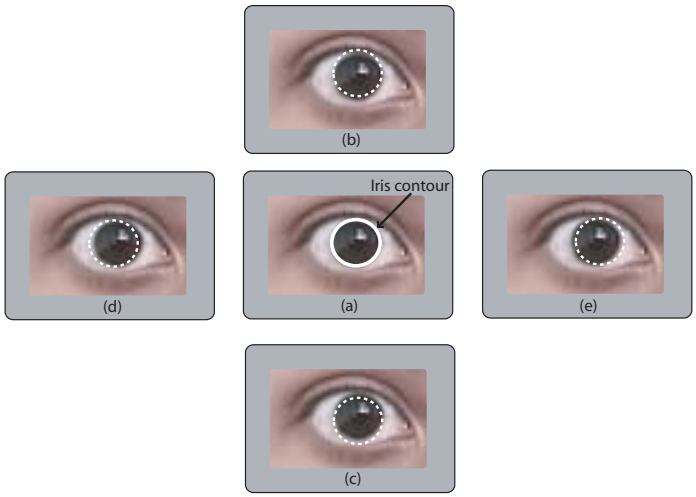

Figure 2.28: Examples of shift tolerance (2 pixels) used in precision evaluation of the iris segmentation. (a) iris contour, (b) shift towards the top, (c) shift towards the bottom, (d) shift towards the left, (e) shift towards the right.

	frames from HCE	frames from HF database
rates	8.89	8.88

Table 2.1: False detection rates (in %) of the iris segmentation on the HCE and HF databases.

from a maximum value of 16% until a minimum value of 8% relatively to the iris diameter (error is equal to 2/D, D being the iris diameter).

The numerical values of the false detection rates are summarized in Table 2.1. Similar rates are obtained on the two databases. The detection rates with a 2 pixels precision are then approximately 91%.

The precision of the iris segmentation results has also been proved to be sufficiently high when used in applications requiring gaze direction estimation as described in the Appendix (see chapter 6).

2.3.2 Limits of the iris segmentation process

The iris segmentation process presents two limitations: minimal size of the face and maximum pan and roll angles.

Based on the maximization of the luminance gradient around a semi-circle, tested on HCE (all the frames), HF (all the frames), YF (40 selected frames), CKE (144 selected frames) and FF (60 selected frames) the minimal size required for the iris segmentation corresponds to one iris of size 5 pixels. Figure 2.29 shows an example of segmentation process with respectively $R = 5$ (good segmentation) and $R = 4$ pixels (segmentation error).

Moreover in the case of pan and roll head motion the results of the segmentation process are accurate as long as the irises are always in the corresponding bounding boxes. As described in the section 2.2.2 each of the right and left irises are searched respectively in the upper left square and upper right square of the face bounding box. These relations depend on the

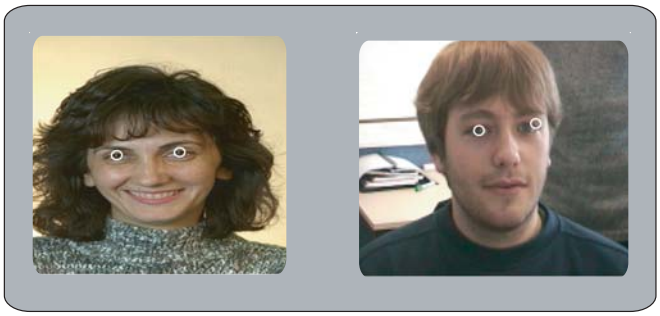

Figure 2.29: Examples of iris segmentation results in the case of its size limit. Left: $R = 5$ (example of FF database), right: $R = 4$ (example of HCE database).

algorithm used for face detection. If we change the face detector, a learning step is necessary to adapt the dimensions of the eye bounding box with respect to the dimensions of the face bounding box. The maximum roll and pan required for a good segmentation result have been measured on the sequences of HF database where the head was rotated according to these two angles. Figure 2.30 shows the horizontal and vertical limits of the head rotation to have good iris segmentation results.

Figure 2.30: Examples of iris segmentation results in the case of roll (left) and pan head motion limits. First row frames from the HCE database, second row frames from the FF database.

2.4 Eyes and eyebrows segmentation

Based on the morphological constraints defined in section 1.2.1, Figure 2.31 gives an example of eyes and eyebrows bounding boxes (solid lines for the eyebrows and dashed lines for the eyes) according to the iris detection. Once the bounding boxes of each facial feature detected, the segmentation approach can be divided into three steps:

- Choice of the model,
- Search of characteristic points and model initialization,
- Model fitting.

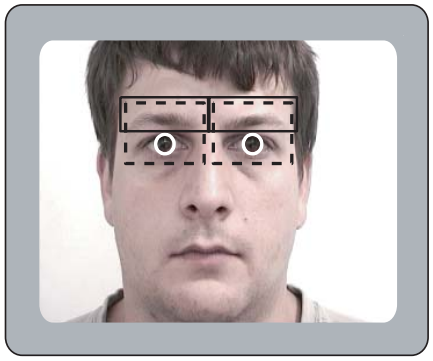

Figure 2.31: Eyes and eyebrows bounding boxes.

2.4.1 Parametric models for eyes and eyebrows

The most common models for eyes and eyebrows boundaries are made of parabolas for the upper and lower eyelids (see [Yui92], [Tia00], [Kap98], [Bot02] and Figure 2.32 middle) and two broken lines defined by three points (both corners and a middle point) for the eyebrows (see [Tia00] and Figure 2.32 left). A fine study on the YF, FF, HCE, CKE databases shows that a parabola is sufficient to model the contour of the lower eyelid, but that the contour of the upper eyelid does not always present a vertical symmetry. Then, a full parabolic model is not always appropriate for the eyes modeling (Figure 2.32 middle). In order to be able to deal with possible different eyelids shapes, Bezier curves based on three control points are used to model the upper eyelid contour (see Figure 2.32 right).

In the same way the study of the eyebrows shape shows that two lines is a bad representation for the eyebrows boundaries (see Figure 2.32 left). Almost due to its great deformation ability under different facial expressions it requires a most flexible model. We choose to model eyebrows by a Bezier curve with three control points. Figure 2.33 right presents the full chosen model for eyes and eyebrows. It is composed of a Bezier curve defined by the three control points $P1$, $P2$, $P3$ for the upper eyelid; of a parabola defined by the points $P1$, $P2$, $P4$ for the lower eyelid; and of a Bezier curve defined by the three control points $P5$, $P6$ and $P7$ for the eyebrows.

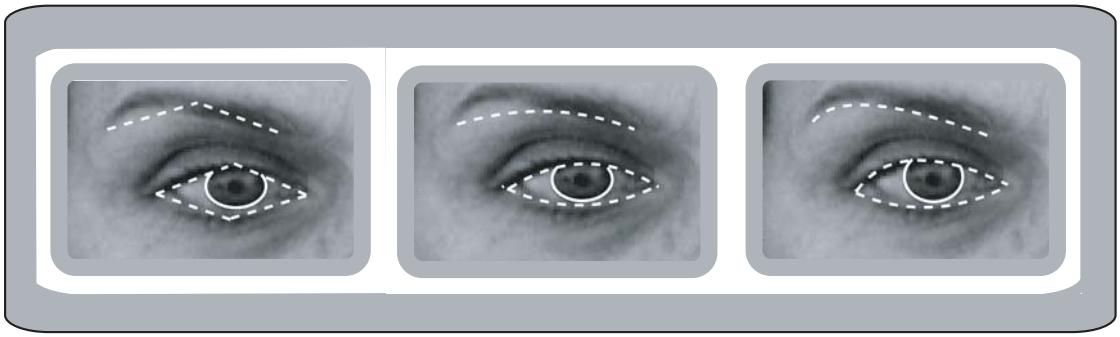

Figure 2.32: Left: eye and eyebrow models with two lines; middle: eye and eyebrow models with two parabolas; right: eye model with a Bezier curve for the upper boundary and a parabola for the lower boundary and eyebrow model with a Bezier curve.

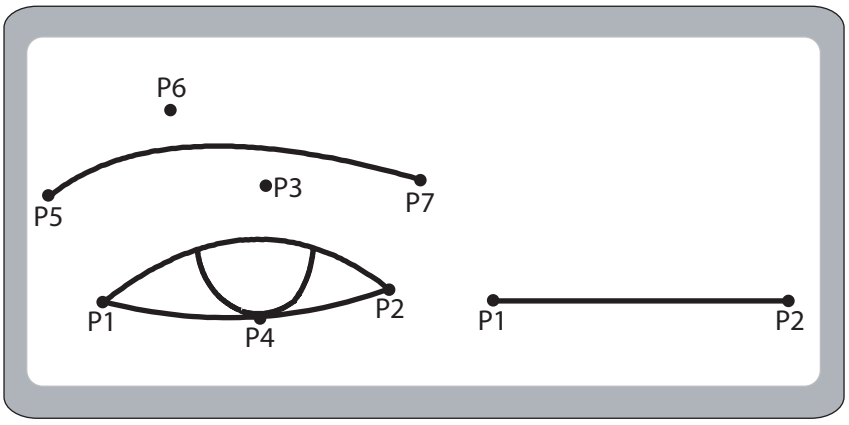

Figure 2.33: Models of eyes and eyebrows and key points.

2.4.1.1 Bezier curves

Originally developed by Pierre Bezier in the 1970's, Bezier curves are well suited to graphics applications (PostScript, Vectorial drawing model, main component of Adobe Illustrator). A Bezier curve is a parametric curve defined by three or four control points (Figure 2.34, Figure 2.35). It has some interesting properties like to be joined together to form smooth (continuous) shapes and they can fold over on themselves. Two examples of Bezier curves according to the spatial configuration of the control points are presented in Figure 2.35. Let's take $P1$, $P2$ and $P3$, three points of the image plane which represent the control points associated to the curve to be defined. In the triangle P1P2P3, we compute the points:

$$M \text{ the barycentre of } P1(1-\lambda) \text{ and } P3(\lambda);$$
$$N \text{ the barycentre of } P2(1-\lambda) \text{ and } P3(\lambda);$$
$$K \text{ the barycentre of } M(1-\lambda) \text{ and } N(\lambda);$$

for a value of λ in the interval $[0, 1]$, λ is the Bezier curve parameter.

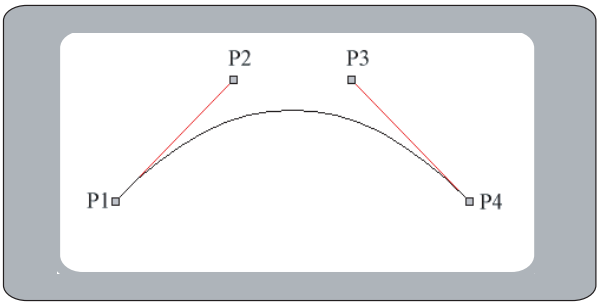

Figure 2.34: Bezier curves with four control points.

By moving the point $P3$, it is possible to obtain different Bezier curves which cross by the point K (see Figure 2.35). The coordinates (x, y) of each point of the parametric Bezier curve

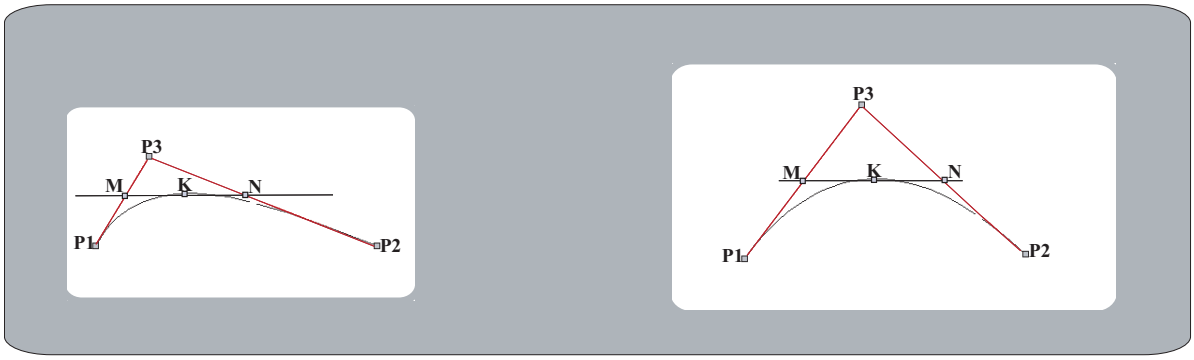

Figure 2.35: A Bezier curve with three control points.

defined by the control points $P1(a_1, b_1)$, $P2(a_2, b_2)$, $P3(a_3, b_3)$ and the parameter λ ($\in [0, 1]$) are obtained by:

$$\begin{cases} x &= (1-\lambda)^2 a_1 + 2\lambda(1-\lambda)(a_3 - a_1) + \lambda^2 a_2 \\ y &= (1-\lambda)^2 b_1 + 2\lambda(1-\lambda)(b_3 - b_1) + \lambda^2 b_2 \end{cases} \tag{2.9}$$

2.4.2 Characteristic points extraction and models initialization

Once the eyes and the eyebrows models have been defined, the next step consists in initializing them. The first stage consists in extracting some characteristic points to be related to each considered facial feature model.

2.4.2.1 Case of eyes

As described in section 1.2.1 according to the position of the iris, the eye search area can be defined (see Figure 2.31). The considered key points for eyes consist in finding in each eye bounding box the three control points of the Bezier curve for the upper contour of the eyelids (P_1, P_2 and P_3) and the three points of the parabolic curve of the lower contour of the eyelids (P_1, P_2 and P_4).

The eyes corners $P1$ and $P2$ are the frontiers between the eye white and the skin (see Figure 2.36) so they are points of local maximum of luminance gradient. Moreover, contours of eyelids correspond to a contour of maximum luminance gradient (frontiers between the eye white and the lashes and between the skin and the lashes). Based on these observations the eyes corners detection consists in tracking the points of maximum luminance gradient towards both the right-hand side and the left-hand side of the detected iris. The two initial tracking points $X1$ and $X2$ are defined as follow: they are pixels of maximum gradient of luminance located under the iris on a vertical line close to the iris circle as described in Figure 2.37.

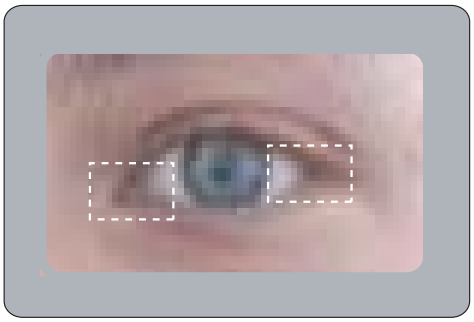

Figure 2.36: Eyes corners characteristics.

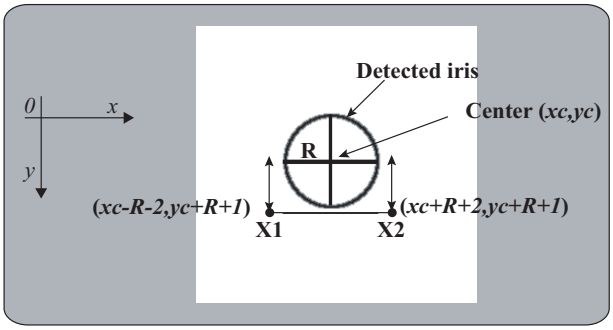

Figure 2.37: Initialization of the tracking points for eyes corners detection.

From each initial point X_1 a tracking process of pixels with the highest value of luminance gradient to the left direction yields to the detection of the first corner $C1$ (see Figure 2.38 top left). For each corner the tracking is delimited by the eyes bounding boxes. The initial point $X1$ is always located below the corner $C1$ (see Figure 2.38 top left). Then at each position of X_1 only the three neighbors (black) points ($\begin{matrix} \bullet & \bullet \\ \bullet & X1 \end{matrix}$) located on the left and above $X1$ are tested. At each step, the luminance gradient of the tested points is evaluated and the point with the maximum gradient of luminance is chosen. The tracking stops when the luminance gradient becomes negative (see example of Figure 2.38 bottom) since a skin pixel is clearer than the eye corner pixel. The resulting curve between $X1$ and $C1$ is made of pixels with local maximum of luminance gradient.

A similar tracking process to the right direction yields to the detection of the second corner $C2$ (see Figure 2.38 top). The neighborhood matrix is changed to $\begin{pmatrix} \bullet & \bullet \\ X2 & \bullet \end{pmatrix}$. Figure 2.39 shows the result of the eyes corners detection. The points $P1$ and $P2$ of the eye model

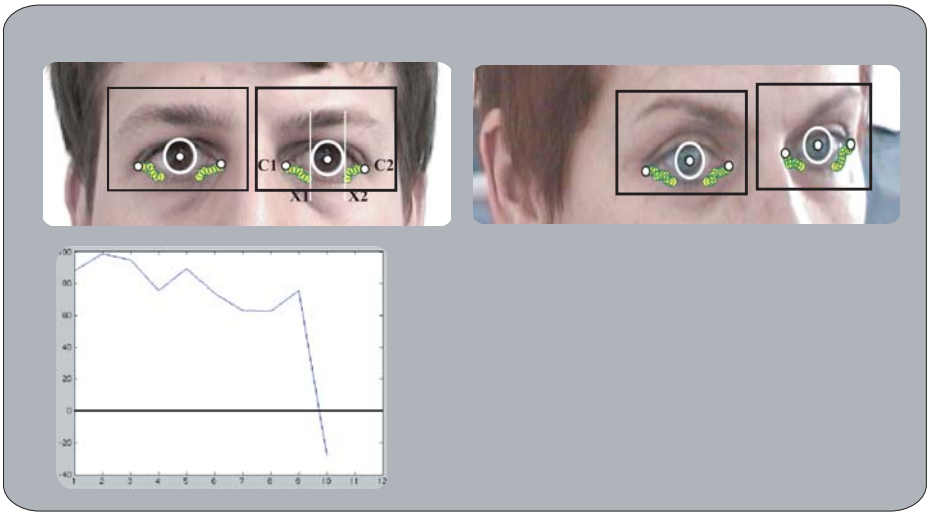

Figure 2.38: (a): Tracking process for the detection of eye corner; (b): luminance gradient evolution along the X_1C_1 curve.

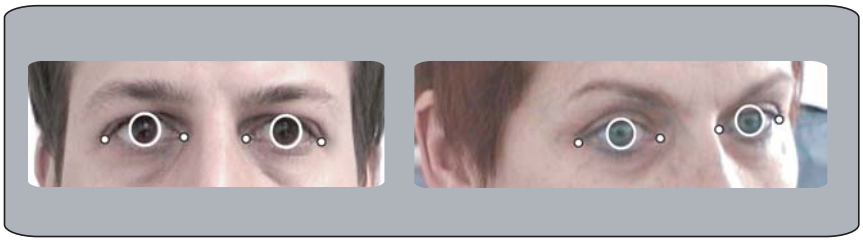

Figure 2.39: Results of eyes corners detection.

(see Figure 2.33) are associated to the two detected corners $C1$ and $C2$. Point $P4$ of the lower parabola is vertically aligned with the lowest point of the detected semi-circle of the iris (coordinates (xc, $yc + R + 1$) with R the iris radius); finally, point $P3$, associated to the upper Bezier curve fits the iris center point. Figure 2.40 presents an example of the automatic initialization of the eye model.

2.4.2.2 Case of eyebrows

The positions and dimensions (Nx, Ny) of each eyebrow search area are defined according to the irises positions (see Figure 2.31 and section 1.2.1). For the eyebrows corners detection the vertical and horizontal projections of the luminance intensity are used. Then the abscissas X_5 and X_7 of the corners $P5$ and $P7$ (see Figure 2.41) in each eyebrow bounding box correspond

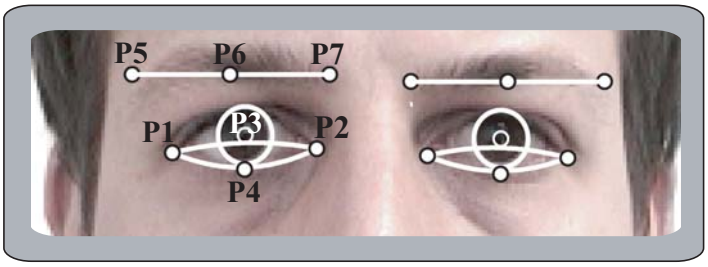

Figure 2.40: Eyes and eyebrows models initialization.

to the left and right zero crossings of the derivative of the quantity:

$$H(x) = \sum_{y=1}^{N_y} [255 - I(x, y)] \tag{2.10}$$

The ordinates Y_5 and Y_7 (with $Y_5 = Y_7$) (see Figure 2.41) correspond to the maximum of the quantity:

$$V(y) = \sum_{x=1}^{N_x} [255 - I(x, y)] \tag{2.11}$$

The third control point $P6$ is computed using the detected positions of $P5$ and $P7$ using the following relations:

$$X_6 = (X_5 + X_7)/2 \tag{2.12}$$

$$Y_6 = Y_7 \tag{2.13}$$

Results of eyebrows corners detection are shown in Figure 2.42. Once the three key points of the Bezier curve detected, the eyebrows model can be initialized (see Figure 2.40).

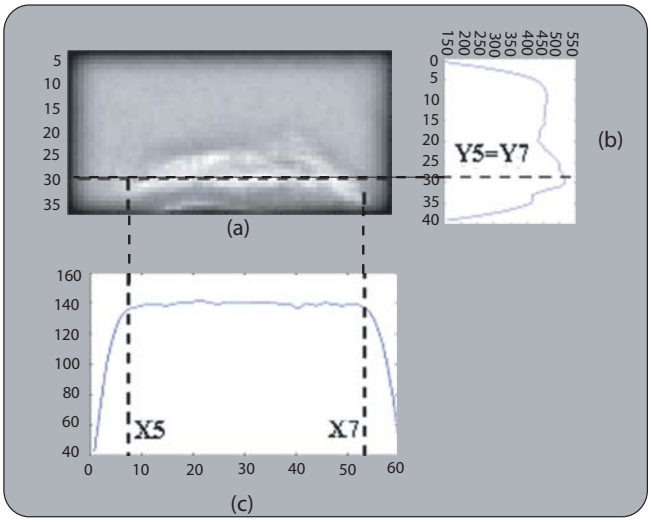

Figure 2.41: From top to bottom and from left to right: video inverse of eyebrow luminance; vertical projection V(y); horizontal projection H(x).

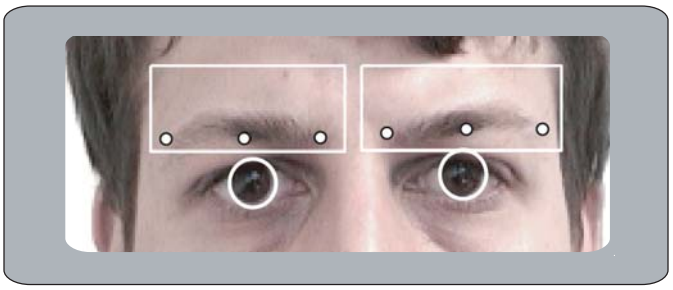

Figure 2.42: Results of the eyebrows key points detection.

2.4.2.3 Eyes and eyebrows key points detection in the case of rotated face

The proposed method is able to detect eyes and eyebrows in the case of rotated face (Figure 2.43). A whole set of geometrical transformations (translation and rotation) has to be performed to correct the detection of eyes and eyebrows key points in the case of head rotation.

Based on the position of the two detected irises the inclination angle θ of the face is computed. θ is the angle between the line joining the centers of the two detected irises and the horizontal (see Figure 2.43) and is equal to :

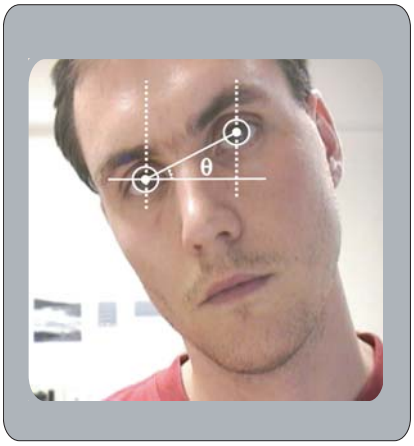

Figure 2.43: Angle of face inclination.

$$\theta = \arctan(y_2 - y_1, x_2 - x_1); \tag{2.14}$$

with (x_1, y_1) and (x_2, y_2) the coordinates of the detected centers of the left and right irises.

For each detected iris, θ is automatically computed and according to its value an inverse rotation is applied to the face to come back to an horizontal position. The eyes and eyebrows key points detection is carried out on the horizontal face position. A reverse transformation gives the position of these points on the initial tilted face. An illustration of this process for eyes corners is given in Figure 2.44.

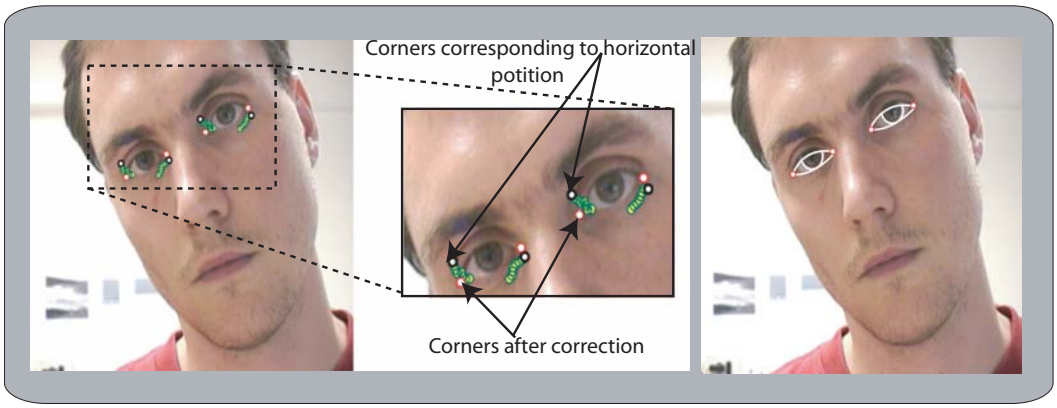

Figure 2.44: Eyes corners detection in the case of a rotated face. left: black points correspond to the detected corners in the horizontal face; right: red (clear in gray level) points correspond to the corrected corners.

Once the eyes corners have been detected, the morphological constraints are used to validate their positions according to the irises centers (see section 1.2.4). If the positions of the detected corner do not check these constraints, another point is computed by default using the result of the section 1.2.4. However, this kind of problems appear sometimes in the case of dark frames or frames exposed to a too strong source of illumination.

2.4.3 Characteristic points extraction improvement with a tracking process

Tested on HCE, CKE, FF and YF databases the segmentation results described before gives good results for eyes and eyebrows key points detection in the case of *Neutral* state and low facial expression intensity (FF database). However, in the case of HCE database and CKE database some false detections appear in the case of higher intensity of the facial expressions (see Figure 2.45), especially for the eyebrows key points detection which are more sensitive to facial expressions (eyebrows undergo more deformations than the eyes in the case of facial expressions). Figure 2.45 shows two detection errors of eyebrows key points in the case of *Disgust* and *Surprise*. These errors are due to the fact that the increase of the facial expression intensity leads to facial features deformations and then the defined morphological constraints are not always verified. For example, sometimes the eyebrows bounding boxes do not contain the eyebrows and sometimes contain noisy information. For example in the case of Figure 2.45 left part of the eyebrows is outside their bounding boxes while in Figure 2.45 right the eyebrows bounding boxes contain the upper eyelids and part of the eyebrows is outside the bounding boxes.

However we can observe that in the case of facial features deformations the key points move progressively from one frame to the next one. Then to overcome this limitation a tracking process is added.

2.4.3.1 Lucas-Kanade algorithm

The tracking process of eyes and eyebrows corners is based on the *optical flow* algorithm developed by Lucas and Kanade in [Luc81]. In this method, the fundamental hypothesis is

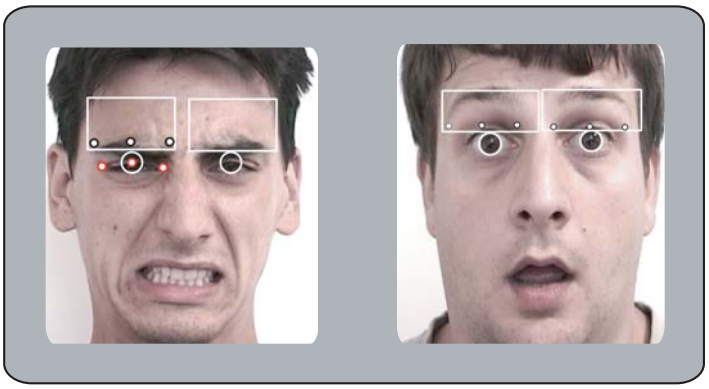

Figure 2.45: Examples of segmentation errors of eyes and eyebrows key points.

that the neighborhood of the tracked point in the frame I_t is found in the following frame I_{t+1} following a simple translation:

$$I_t(x - d(x)) = I_{t+1}(x) \tag{2.15}$$

where $I_t(x)$ and $I_{t+1}(x)$ are the luminance values in two temporally adjacent frames at time t and at time $t+1$ and $d(x)$ is a vector of displacement of the pixel of coordinate x (x a vector). Let us consider a neighborhood R of size nxn in the frame I_t. The aim of the process is to find in I_{t+1} the most similar area to R. For this, it is necessary to minimize a cost function equal to the sum of the squared differences inter-pixels:

$$\varepsilon(d(x)) = \sum_{x \in R} [I_t(x - d(x)) - I_{t+1}(x)]^2 w(x) \tag{2.16}$$

where $w(x)$ is weight function.

Generally, $w(x)$ is constant and is equal to 1. But it can also take a Gaussian form if we want to give more importance to the center of the window. The minimization of the function ε is carried out in an iterative way. One notes $d^i(x)$ the value of the displacement computed after the iteration i. The final displacement $d^{i+1}(x)$ at iteration $i+1$ is expressed:

$$d^{i+1}(x) = d^i(x) + \Delta d^i(x) \tag{2.17}$$

where $\Delta d^i(x)$ is the incremental displacement to be determined with a sub-pixel precision. We make the hypothesis that the considered neighborhood does not undergo any deformation. Consequently the value of displacement is the same one for all the pixels of R. equation 2.17 can thus be rewritten as:

$$d^{i+1} = d^i + \Delta d^i \tag{2.18}$$

At the beginning of the process, $d^0 = [00]^T$. Lucas-Kanade algorithm provides correct tracking results from one frame to the next one. But after several consecutive frames small errors are progressively accumulated and the obtained key points position tends to diverge from the real positions. Then the eyes and eyebrows key points detection is made by combining the Lucas-Kanade tracking process and the static extraction as explained in the following section.

2.4.3.2 Case of eyes

Lucas-Kanade algorithm is introduced in the eyes key points detection ($P1$, $P2$, $P3$ and $P4$). It has been tested on the HCE database (tracking needs a video sequence which is not the case of the other databases).

For the eyes corners $P1$ and $P2$ the tracking results are very closed to those of the automatic detection Indeed, we do not have any means to identify the most accurate positions between them. Based on our visual expertise we observe that even a human observer can not decide which are the most accurate positions between the tracked (see Figure 2.46 black positions) and the detected positions (see Figure 2.46 clear positions). For this reason we choose to take into account both results. At each time, the eyes corners positions correspond to the mean position between the position obtained by the tracking process and the automatic segmentation process (see Figure 2.46).

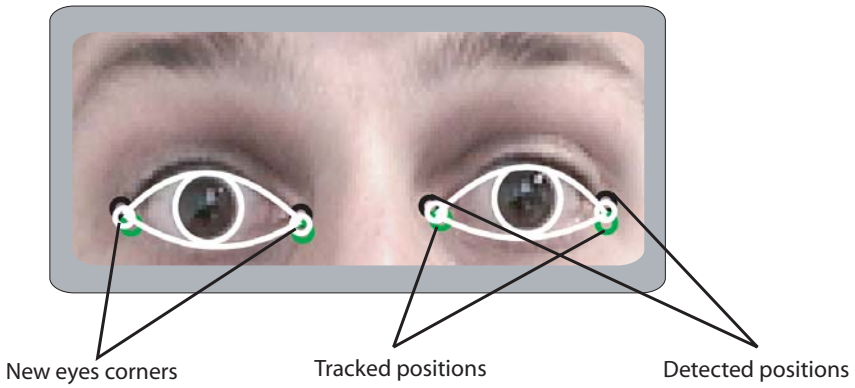

Figure 2.46: New eyes corners position (white) according to the tracked and detected positions (clear and black).

2.4.3.3 Case of eyebrows

In opposition to the eyes contours, facial expressions induce non rigid deformations of eyebrows. This leads to some errors in the eyebrows bounding boxes detection and then in the detection of the eyebrows key points (see Figure 2.45). As described in the section 2.4.2.2, the eyebrows corners detection is based on the horizontal and vertical sum of the luminance intensity of each selected eyebrows bounding box (see Figure 2.41). Then a false detection of the eyebrows bounding box leads to a change in the luminance distribution of the pixels in this area and consequently to a false detection of the eyebrows key points (see Figure 2.45).

As described in section 1.2.1 the horizontal position of the upper limit of the eyebrows bounding boxes noted y_{sup} is defined by $y_{sup} = y_c - R$ (see Figure 2.47). This limit is accurate in the case of a *Neutral* face but leads to some errors in the case of facial deformations. Figure 2.45 right shows the case where the inner eyebrows corners are lowered. Based only on the iris position the computed y_{sup} leads to false detection in the eyebrows bounding box detection. In this case the bounding boxes contain the upper eyelids and only a part of the eyebrows. Figure 2.45 left presents the case where the inner eyebrows are raised. In this case, the detected eyebrows bounding boxes may not contain it. Then in the two cases the eyebrows key points detection leads to detection errors. To handle the facial deformations, the eyebrows

bounding boxes must be sufficiently large to entirely surround the eyebrows but sufficiently small not to take into account noisy information (for example eyelids). Hence at each frame (apart from the first one) the horizontal position of the upper limit of the eyebrows bounding boxes y_{sup} is defined by the horizontal position of the third control point of the Bezier curve corresponding to the last fitted upper eyelid (at time $t-1$). In this way at each time the eyebrows bounding boxes are defined according to the eyebrows motion in order to be sure that their bounding boxes surround the eyebrows when these latter are raised (for example in the case of *Disgust* expression Figure 2.48 left). However this solution is not sufficient to solve the errors occurring when the eyebrows are lowered (for example *Surprise* expression in Figure 2.48 right). To deal with this motion (lowered eyebrows) the tracking process is combined with the detection.

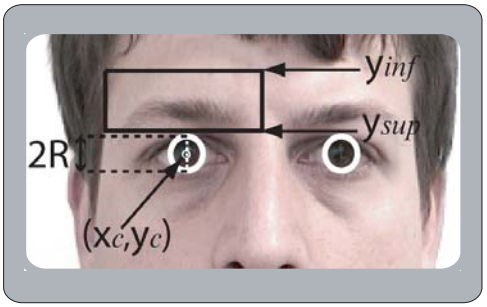

Figure 2.47: Eyebrows bounding boxes selection.

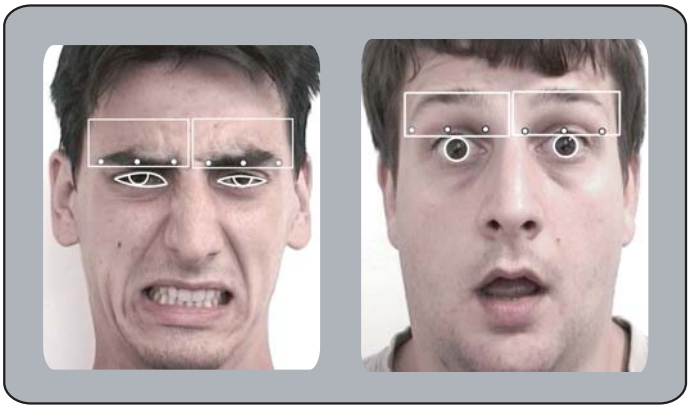

Figure 2.48: Left: correct detection of eyebrows key point; right: false detection of eyebrows bounding boxes and, then, resulting false detection of the eyebrows key points.

At each time, the fusion process consists in computing the static and the tracked eyebrows key points and then in respectively computing their relative position according to the one obtained in the previous frame. The one which is the closest to the last corresponding position is then chosen. Figure 2.49 presents an example of this selection process. Figure 2.57 presents the segmentation results of eyes and eyebrows before and after the intoduction of the dynamic information.

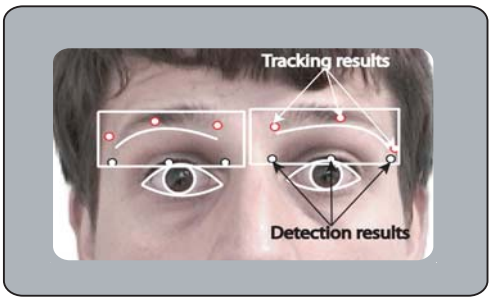

Figure 2.49: Tracked (red points (clear in gray level)) and the detected (black (dark in gray level)) eyebrows key points. The selected eyebrows key points in this case are the tracked ones.

2.4.4 Eyes and eyebrows key points precision

In order to have a quantitative evaluation of the eyes and eyebrows key points detection, a comparison with a manual ground truth has been done. To realize this comparison, 1680 frames coming from 8 different subjects from the HCE database have been manually labeled and also automatically segmented using the proposed algorithm.

For each frame i, the reference points k is noted $P_{k,ref}(i)$. The corresponding detected point by our algorithm is noted $P_{k,detect}(i)$. The average error for each characteristic point is noted $\epsilon_{k,detected}$ and has been obtained as:

$$\epsilon_{k,detected} = \frac{1}{NB}\sum_{i=1}^{NB}\frac{|\overrightarrow{P_{k,ref}(i)P_{k,detect}(i)}|}{|\overrightarrow{P_{A,ref}(i)P_{B,ref}(i)}|} \tag{2.19}$$

where NB corresponds to the number of labeled frames. The detection errors are reported relatively to eyes width ($P_A = P1$ and $P_B = P2$) and eyebrows width respectively ($P_A = P5$ and $P_B = P7$).

Moreover in order to evaluate the error of the human expert on the detection of the key points, on 15 randomly selected frames from the HCE database the 6 key points are manually selected ten times. As the manual detection is not perfect, it exists a dispersion of the position of the points which is computed as:

$$\epsilon_{k,human} = \frac{1}{NB}\sum_{i=1}^{NB}(\frac{1}{M}\sum_{m=1}^{M}\frac{|\overrightarrow{P_{k,ref}(i)P_{k,human}(i,m)}|}{|\overrightarrow{P_{A,ref}(i)P_{B,ref}(i)}|}) \tag{2.20}$$

where M corresponds to the session number of the manual detection (M=10). $P_{k,human}(i)$ is the characteristic point k manual detection in the session m of the frame i. Moreover, $P_{k,ref}$ is the average of the manually detection of the point k on the frame i:

$$P_{k,ref}(i) = \frac{1}{M}\sum_{m=1}^{M}P_{k,human}(i,m) \tag{2.21}$$

The numerical accuracy values are summarized in Table 2.2. As it can be observed, even with a manual labeling, the relative error is not null (cf. Table 2.2).

	$P1$	$P2$	$P5$	$P7$
auto	4.6	3.0	7.8	4.90
hand	2.0	1.9	1.6	1.4

Table 2.2: Relative errors (in %) after the automatic extraction (auto) and the manual extraction (hand) of the key points.

2.4.5 Models fitting

In order to fit the initial models to the contours of the image, a deformation criterion is required. Our approach is based on luminance gradient maximization along each candidate curve for eyes and eyebrows.

2.4.5.1 Case of eyes

After testing on the whole of HCE and HF databases, we conclude that the initial curve for the contour of the lower eyelid is satisfactory and that it is not necessary to move it (see Figure 2.40). However, it is necessary to find an evolution criterion for the upper curve in order to adapt it to the shape of the upper eyelid.

As described before, the eye boundary is a curve made of pixels with a maximum gradient of luminance. Using this information the best curve is selected by the maximization of the normalized flow of luminance gradient $NFLG_t$ (same expression as equation 2.6) through the upper eyelid contour. Then, the third control point $P3$ of the upper Bezier curve initialized at the iris center position has to be moved upwards (the control points $P1$ and $P2$ being fixed) in order to find the curve which maximizes the $NFLG_t$. $P3$ is moved inside the rectangle area described in Figure 2.50 right eye. To define this search area, the analysis of the different positions of this point after the fitting of the upper eyelid contour is made on the HCE and HF databases. This study shows that in the case of *Neutral* faces this area is always delimited at the top by the horizontal position of the third key point $P6$ associated to the eyebrow (see section 2.4.2.2). At the bottom, by the center of the iris and at the left and right side of the iris by $2R$ (see Figure 2.50).

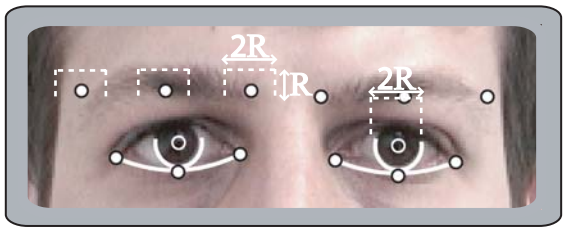

Figure 2.50: left: scanning area of the three eyebrows key points $P5$, $P6$ and $P7$; right: scanning area of the eyelid control point $P3$.

However, tested in the case of facial expressions, the defined scanning area of $P3$ leads to lowest accuracy results in the apex of the expression. For example, when the degree of eyes opening is low and the eyebrows corners are raised, the limits between the upper eyelids and the skin is very dark and sometimes the eyebrows come close to the eyelids (see Figure 2.51 left). For this, from its initial position (the center of iris position) $P3$ must take into account possible facial deformations during the scanning process inside the eye bounding box and then its scanning area has to be redefined at each time. To do this, at each time t (new frame) the scanning area of the key point $P3$ is horizontally delimited by the last position of $P6$ (its horizontal position at time $t-1$ after a fitting process of the eyebrows model, see Figure 2.51). So in the first frame, the scanning area of $P3$ is delimited by the horizontal position of the middle key point of the eyebrow model $P6$ before the curve fitting process and in the following of the sequence, by the last position of this point obtained after the fitting process of the eyebrow model. The bottom, the left and right side do not change. At each position

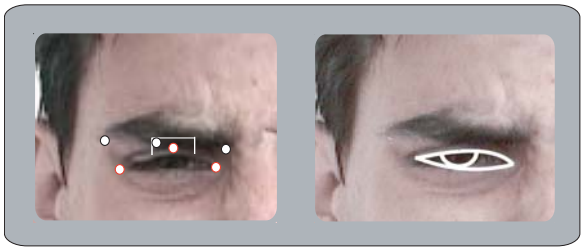

Figure 2.51: Left: scanning area for $P3$; right: corresponding segmentation result of the eye.

of $P3$ inside this search area is associated a new Bezier curve. Figure 2.52 presents several tested curves (white curves). For each curve, the normalized sum of luminance gradient is computed. Figure 2.52 right shows the evolution of the normalized gradient of luminance $NFLG_t$ for all the tested curves during the scanning process of the control point $P3$ inside the search area. The one with a maximum of luminance gradient is chosen to be the upper eyelid (red curve (clear in gray level) in Figure 2.52 left).

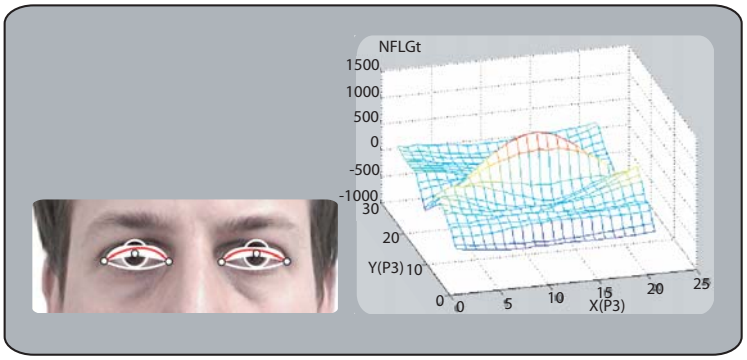

Figure 2.52: Left: maximization of the $NFLG_t$ and several tested curves (curves in white) for the eyes and the selected one (curve in red (clear in gray level)); right: evolution of the $NFLG_t$ when $P3$ is scanning the search area; the maximum of the $NFLG_t$ gives the Bezier curve which fits the upper contour of each eye (red curve in left).

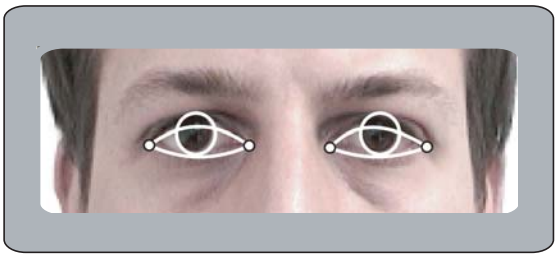

Figure 2.53: Eyes segmentation results.

The detection of the iris is based on the research of the semi circle with maximum of luminance gradient. Each selected semi-circle is completed thereafter by symmetry (see Figure 2.53). However, the eye is not always widely open and the higher part of the iris is sometimes partially hidden by the top eyelid. Thus to obtain a better detection of the iris in the open state, it is necessary to correct the result of the completed symmetry. The correction of the irises shapes consists in eliminating the superfluous part of the detected irises. Using the upper eyelids contour, it is possible to compute the intersection between the Bezier curve and the curve of the iris. Figure 2.54 shows examples of irises correction. On the left the shape of the initial irises are presented and on the right their corrected contour.

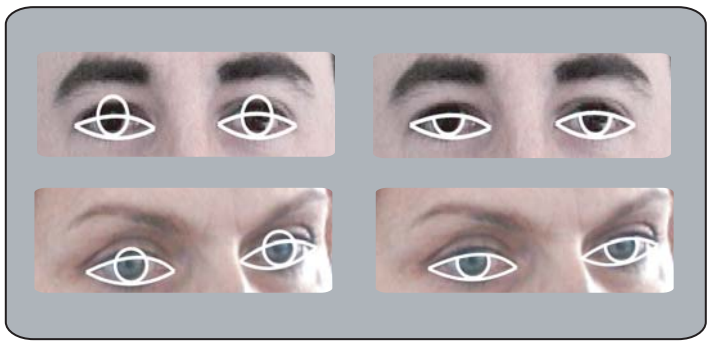

Figure 2.54: Results of iris correction; left: before the correction; right: after the correction.

2.4.5.2 Case of eyebrows

In the case of eyebrows the three control points $P5$, $P6$ and $P7$ are moved successively to fit the eyebrows contour. This is due to the fact that the automatic detection of the positions of $P6$ and $P7$ is less precise than the one of $P1$ and $P2$. Similarly to the case of eyes the analysis of the positions of the three key points of the eyebrows is made on the HCE and HF databases. This analysis allows to define for each key point a search area inside which it will be moved. It corresponds to the area centered around their initial position with a width of $2R$ and a height of R (see Figure 2.50).

A new Bezier curve is associated at each position of $P5$ $P6$ $P7$ inside their search area. Figure 2.55 present several tested curves (white curves). As in the case of the eyes the curve maximizing the $NFLG_t$ is chosen to be the eyebrows contour (red curve (dark in gray level) in Figure 2.55). Figure 2.56 shows an example of eyes and eyebrows segmentation.

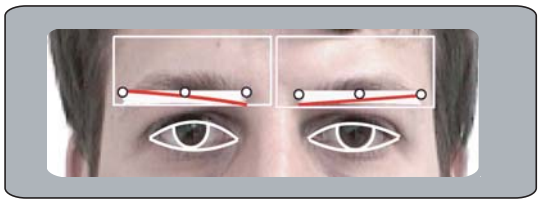

Figure 2.55: Eyebrows model fitting; white curves: intermediate curves; red curve (clear in gray level): selected curve.

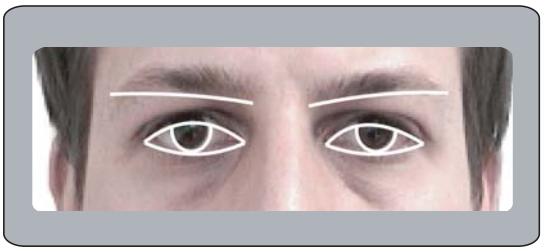

Figure 2.56: Eyes and eyebrows segmentation results.

Figure 2.57 presents the segmentation results of eyes and eyebrows before and after the temporal correction of eyes and eyebrows key points. We can observe that the accuracy of the segmentation results are very comparable in the case of *Neutral* expression (first column) and weakly expressive faces (second column). However we can observe that the segmentation results are better in the case of very expressive faces (third column) after the introduction of the temporal information (second and fourth rows).

Figure 2.57: Eyes and eyebrows segmentation results before (rows 1 and 3) and after (rows 2 and 4) temporal information (tracking process and adaptative scanning area for eyes and eyebrows key points).

2.5 Mouth segmentation

The mouth segmentation is out the scope of our work. For this we use the lips segmentation algorithm developed in [Eve03a].

The mouth segmentation is based on the same principle of eyes and eyebrows segmentation (model choice, model initialization and model fitting). Contrary to the eyes and the eyebrows the luminance information is not sufficient to detect the lips boundaries and the method is also based on chrominance information. In the following we summarize the mouth segmentation approach, more details can be found in [Eve03a].

2.5.1 Mouth model

The model which is made of 5 independent curves is used for mouth segmentation. Each curve describes a part of the lip boundary. Between Q_2 and Q_4, the Cupidon's bow is drawn with a broken line and the other parts of the contour are approximated by 4 cubic polynomial curves γ_i (see Figure 2.58). It is also considered that each cubic has a null derivative at key points Q_2, Q_4 or Q_6.

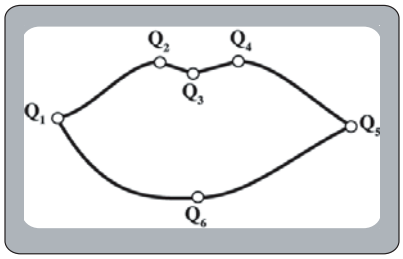

Figure 2.58: Mouth parametric model (see [Eve03a]).

2.5.2 Mouth key points detection

The localization of characteristic points of the mouth uses mixed information of luminance and chrominance and a new kind of snake called *jumping snake*.

Color and luminance information are combined for the lips detection. It works in the RGB color space and it uses the "hybrid edge" $R_{top}(x, y)$ introduced in [Eve02]. It is computed as follows:

$$\overrightarrow{R_{top}(x, y)} = \overrightarrow{\nabla}[h_N(x, y) - I_N(x, y)] \tag{2.22}$$

where $h_N(x, y)$ and $I_N(x, y)$ are respectively the pseudo hue and the luminance of pixel (x, y), normalized between 0 and 1. ∇ is the gradient operator. This hybrid edge exhibits much better the top frontier of the mouth than the classic gradients of luminance or pseudo-hue (see [Eve01]).

To find the upper mouth boundary, a new kind of active contour called "jumping snake" is introduced. Its convergence is a succession of jump and growth phases [Eve03b]. It is initialized with a seed S^0 that can be located quite far away from the final edge (see Figure 2.59 left). The seed is put manually above the mouth and near its vertical symmetry axis.

The snake grows from this seed until it reaches a pre-determined number of points. Then, the seed 'jumps' to a new position that is closer to the final edge. The process stops when the size of the jump is smaller than one pixel (which requires 5 iterations in average). Six principal key points are used (see Figure 2.58): the right and left mouth corners (Q_1 and Q_5), the lower central point (Q_6) and the three points of the Cupidon's bow (Q_2, Q_3 and Q_4). They are used to find the lower central point Q_6. The three upper points are located on the estimated upper lip boundary resulting from the jumping snake algorithm. Q_2 and Q_4 are the highest points on the left and right of the seed. Q_3 is the lowest point of the boundary between Q_2 and Q_4 (see Figure 2.59 right). The points Q_6 is found by analyzing $\nabla_y(h)$, the $1D$ gradient of the pseudo hue along the vertical axis passing by Q_3 (see Figure 2.59 right).

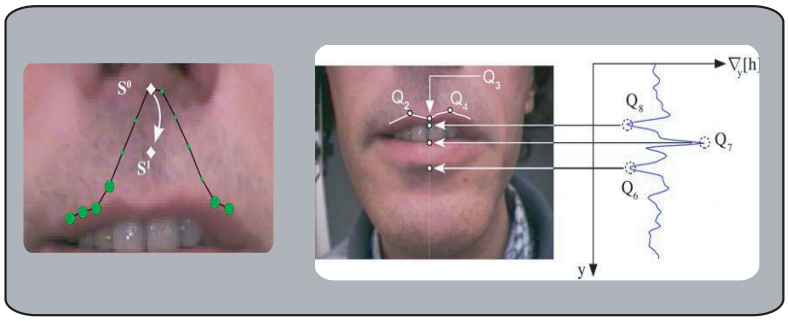

Figure 2.59: Left: jumping snake initialization and seed jump; right: the 3 upper points are found on the estimated upper boundary resulting from the jumping snake algorithm (white line). $Q6$ is below $Q3$, on extrema of $\nabla_y[h]$ (see [Eve03a]).

2.5.3 Model fitting

A cubic curve is uniquely defined if its four parameters are known. Here, each curve passes by, and has a null derivative on points Q_2, Q_4 or Q_6. These considerations bring 2 constraint equations that decrease the number of parameters to be estimated from 4 to 2 for each cubic. So, only two more points of each curve are needed to achieve the fitting. These missing points are chosen in the most reliable parts of the boundary, i.e. near Q_2, Q_4 or Q_6. Upper curves missing points have already been found by the jumping snake. On the other hand, only one point (Q_6) of the lower boundary is known. To get additional lower points, it makes a snake grow from the seed $Q6$. The growth stops after a few points (the white dots in Figure 2.60). Now that there are enough points for each part of the boundary, it should be possible to compute the curves γ_i passing by them to find the mouth corners where these curves intersect.

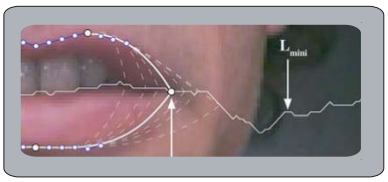

Figure 2.60: The dotted curves are the cubic curves associated to the different tested points along L_{mini} (see [Eve03a]).

In order to make the results more robust, it is also supposed that the corners (Q_1, Q_5) are known. Mouth corners are considered located in dark areas corresponding to the minimum of luminance for each column between the upper and the lower boundary. It gives the line of minima Lmini (see Figure 2.60) and the corners are supposed located on this line. A given (right or left) corner corresponds to a unique couple of upper and lower curves (the dashed curves in Figure 2.60). So, the fitting is achieved by finding the corners that give the best couple of curves (see Figure 2.61).

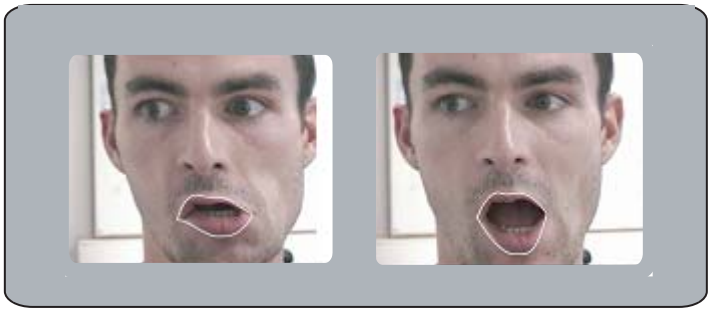

Figure 2.61: Example of mouth segmentation results [Eve03a].

2.6 Face tracking process

In order to track the face bounding box in a video sequence according to the manually extracted face in the first frame, a block matching method is used. It consists in tracking from one frame to the next one the most similar area to a reference one using the Euclidean distance defined by:

$$D_{face} = \sum_{(x,y)\in \text{facezone}} (I_t(x,y) - I_{t+1}(x,y))^2 \tag{2.23}$$

where $I_t(x,y)$ and $I_{t+1}(x,y)$ are the intensities at position (x,y) of the frames respectively at time t and at time $t+1$.

The tracking process have been tested on all the sequences of the HCE database. The results show that our method for face bounding box tracking is very accurate and robust to roll and pan head motion (see Figure 2.62) with the only assumption that there is always a face in each frame.

2.7 Eyes and eyebrows segmentation results

The performances and the limits of the eyes and eyebrows segmentation method are highlighted thanks to the analysis of a great number of results (YF, HF, FF, HCE, CKE databases).

2.7.1 Pertinence of the models

Segmentation results of Figure 2.63, Figure 2.64 and Figure 2.65 show the pertinence of the chosen models for eyes and eyebrows contours. Upper eyelid contours and eyebrow contours

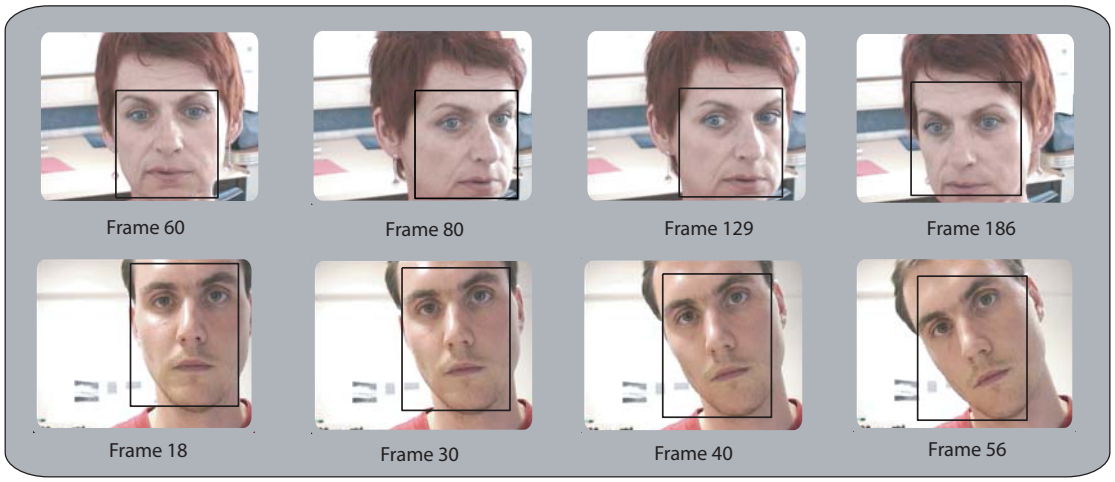

Figure 2.62: Face tracking results in the case of top: pan head rotation; bottom : roll head rotation.

do not always present a vertical symmetry so that parabolas are not well suited for such features (Figure 2.63 image 1 of row 3, Figure 2.65 image 2 of row 1 and image 1 of row 2, Figure 2.64 images 1, 2 and 3). A visual evaluation of the pertinence and flexibility of the chosen models for the eyes and eyebrows contours show also their great ability to deal with facial features deformations. Eyes and eyebrows are sensitive to the facial expressions and undergo strong deformations according to the expression. They can be slackened in the case of *Smile* or *Neutral*, downwards in the case of *Disgust* and towards outside in the case of *Surprise*. Figure 2.63 and Figure 2.65 show some segmentation results with different expressions intensities. These results are related to the chosen models and especially for the eyebrows where the model is able to fit them even in the case of very high intensity of the expressions. The segmentation results on the different databases show also the suitability of the models to the ethnicity variation (Figure 2.63, 2.65 and 2.64).

Figure 2.63: Eyes and Eyebrows segmentation results on HCE database. R =7 (R being iris radius).

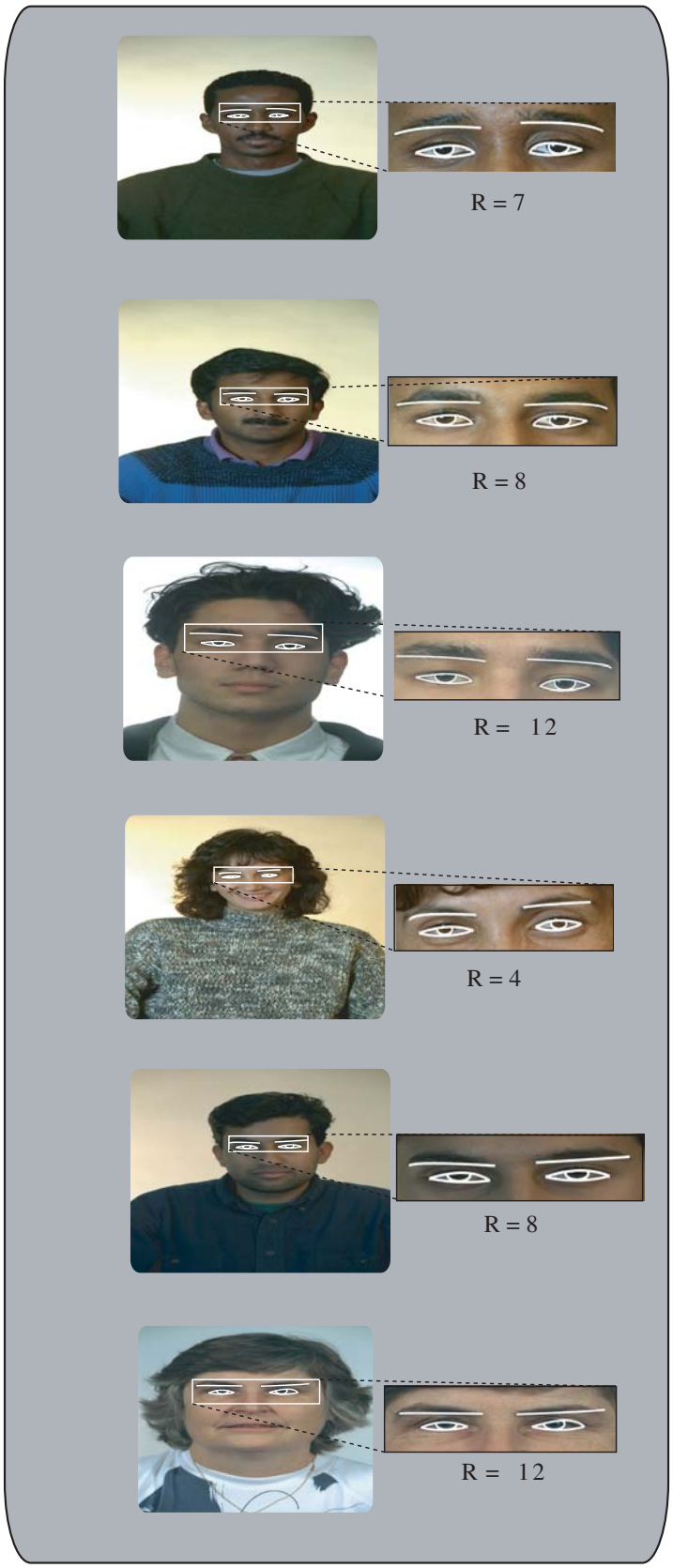

Figure 2.64: Eyes and Eyebrows segmentation results on the FF database (R iris radius).

Figure 2.65: Eyes and Eyebrows segmentation results on the CKE database.

2.7.2 Robustness to spectacles

Figure 2.66 shows also examples of segmentation in the case of spectacles. Segmented contours are of good quality even if the subject is wearing spectacles. Spectacles are not a problem since the first extraction consists in the iris semi circle detection and spectacles are in general bigger than the eyes. As far as the eyebrow segmentation is concerned, it is possible that in some cases, the eyebrow contour coincides with the spectacles one. But in such cases, it is not possible to make the difference even visually.

Figure 2.66: Segmentation results in the case of spectacles on the HCE (first two rows) and the FF database (last row) (R being iris radius).

2.7.3 Robustness to luminance conditions

Figure 2.67 presents the robustness of the proposed method to varying luminance conditions (the left/right part of the face is strongly exposed to a light source) thanks to the retinal

prefiltering. We can also see the accuracy of the segmentation results to bad luminance conditions even with spectacles (see Figure 2.70).

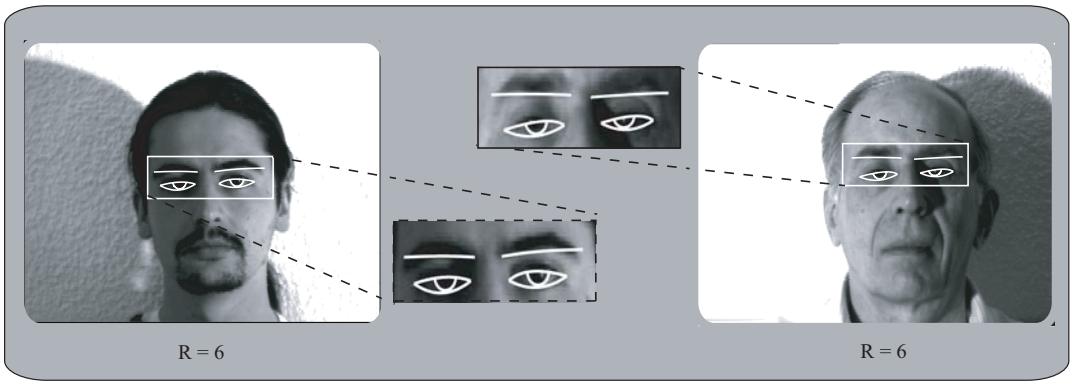

Figure 2.67: Segmentation results in the case of bad luminance conditions. Frames from the YF database (R being iris radius).

2.7.4 Robustness to 2D planar head motion

Figure 2.68: Segmentation results in the case of horizontal and vertical head rotations. First row frames from the HCE database; second row frames from the FF database (R being iris radius).

Figure 2.68 shows the robustness of the proposed method to pan and roll head rotations. It presents the limit of the pan and roll face rotation where the system gives a good segmentation result. Based on the irises detection the system gives a satisfactory segmentation result provided the two irises are accurately detected. Then, similarly to the irises detection, the limits of the eyes and eyebrows segmentation algorithm to head motion corresponds to: roll $\approx 45^o$ and pan $\approx 40^o$.

2.7.5 Limitations of the segmentation process

2.7.5.1 Small face

Based on the segmentation results under various faces sizes (corresponding to iris radius between 4 to 13 pixels) in the whole set of databases, the most accurate results (from a human observer judgment) of the proposed segmentation method is obtained where the face size corresponds to an iris radius R higher than 4 pixels. The lower the eyes and eyebrows size is, the less accurate their segmentation is. These results are due to the low resolution of the eyes and eyebrows which make their key points detection very difficult leading to a false initialization of the eyes model (see Figure 2.69) and then to false segmentation results.

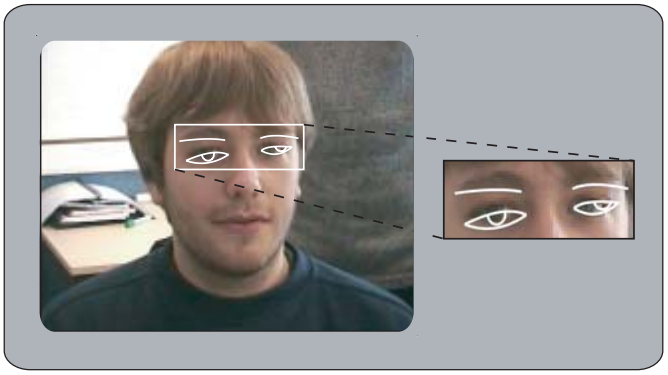

Figure 2.69: Eyes segmentation in the case of small face dimensions corresponding to iris radius $R = 4$.

2.7.5.2 Spectacles with bad luminance conditions

A false detection of eyes and eyebrows can appear in case of very dark frames where the frontier between the eye white and the skin (resp the frontier between the eyebrow contour and the skin) is difficult to distinguish even by human observer and when it is associated with the presence of light reflection on spectacles which realizes a smoothing of the eyelid and eyebrows contours (see Figure 2.70). In this case even if the eyes characteristic key points are accurately detected, inside the associated search area, the maximization of the luminance gradient along the Bezier curves can fail.

2.7.6 Precision of eyes and eyebrows segmentation

Similarly to the evaluation of the iris precision, in order to have a quantitative, it is necessary to have a *ground truth*. In order to obtain a labeled database with a sufficient number of

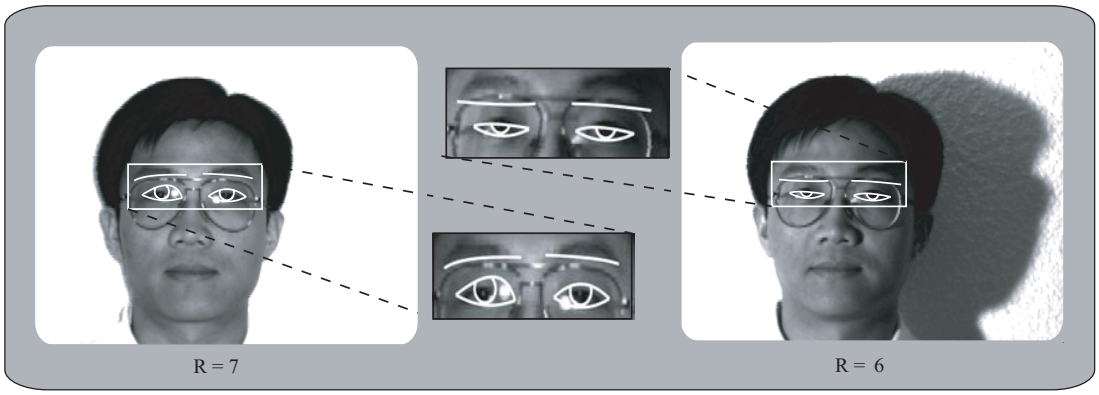

Figure 2.70: Segmentation results in the case of spectacles and under bad luminance conditions (R iris radius).

samples (several hundreds of frames), we should manually detect a great number of points along the contours. However this is an herculean task and as a consequence, we have only evaluated the precision of the eyes and eyebrows control points of the Bezier curves (P1,P2,P3 for the eyes and P5,P6,P7 for the eyebrows) detected by our algorithm. To realize this comparison, the same frames used for the evaluation of the initial key points precision have been used (1680 frames manually labeled on the one hand and automatically segmented on the other hand using the proposed algorithm (see section 2.4.4).

The average error for each characteristic point has been obtained using formula 2.19 and the dispersion of the position of the manually detected points has been computed using formula 2.20 (see section 2.4.4).

The numerical accuracy values are summarized in Table 2.3. The automatic segmentation allows the estimation of the key points P1,P2,P3,P5,P6,P7 with a precision comparable to a manual segmentation.

	$P1$	$P2$	$P3$	$P5$	$P6$	$P7$
auto	4.6	3.0	5.7	4.1	6.9	4.2
hand	2.0	1.9	2.6	1.6	2.0	1.4

Table 2.3: Relative errors (in %) of the eyes and eyebrows key points after the automatic extraction (auto) and the manual extraction (hand) of the key points.

2.7.7 Permanent facial features segmentation

Figure 2.71 gives final segmentation results of eyes, eyebrows and lips on the HCE database. We can observe the robustness to facial expressions leading to great deformations of the facial features.

2.8 Conclusion

We have presented a segmentation method of iris, eyes and eyebrows based on luminance gradient. We have introduced different models: a circle for the iris, a Bezier curve for the upper eyelid, a parabola for the lower eyelid and a Bezier curve for the eyebrow. The chosen models are very flexible and enable the method to be robust to all the types of eyes and eyebrows deformations. Intensive tests on several databases presenting a large range of varying conditions show the accuracy and the robustness of the method to spectacles, luminance conditions, ethnicity and facial features deformations. The method is accurate for face dimensions higher or equal to $90 * 90$ pixels corresponding to an iris radius $R \geq 4$.

The two main limitations of the proposed algorithm are the manual detection of the face in the first frame and the iris radius. A face detector can be used (for example MPT [Mac99] and OpenCV face detector [Ope06]) subject to the adaptation of the eye bounding boxes limits. It could be also possible to solve the problem of the manual estimation of the iris radius by testing several radius values since this one is correlated to the face dimensions.

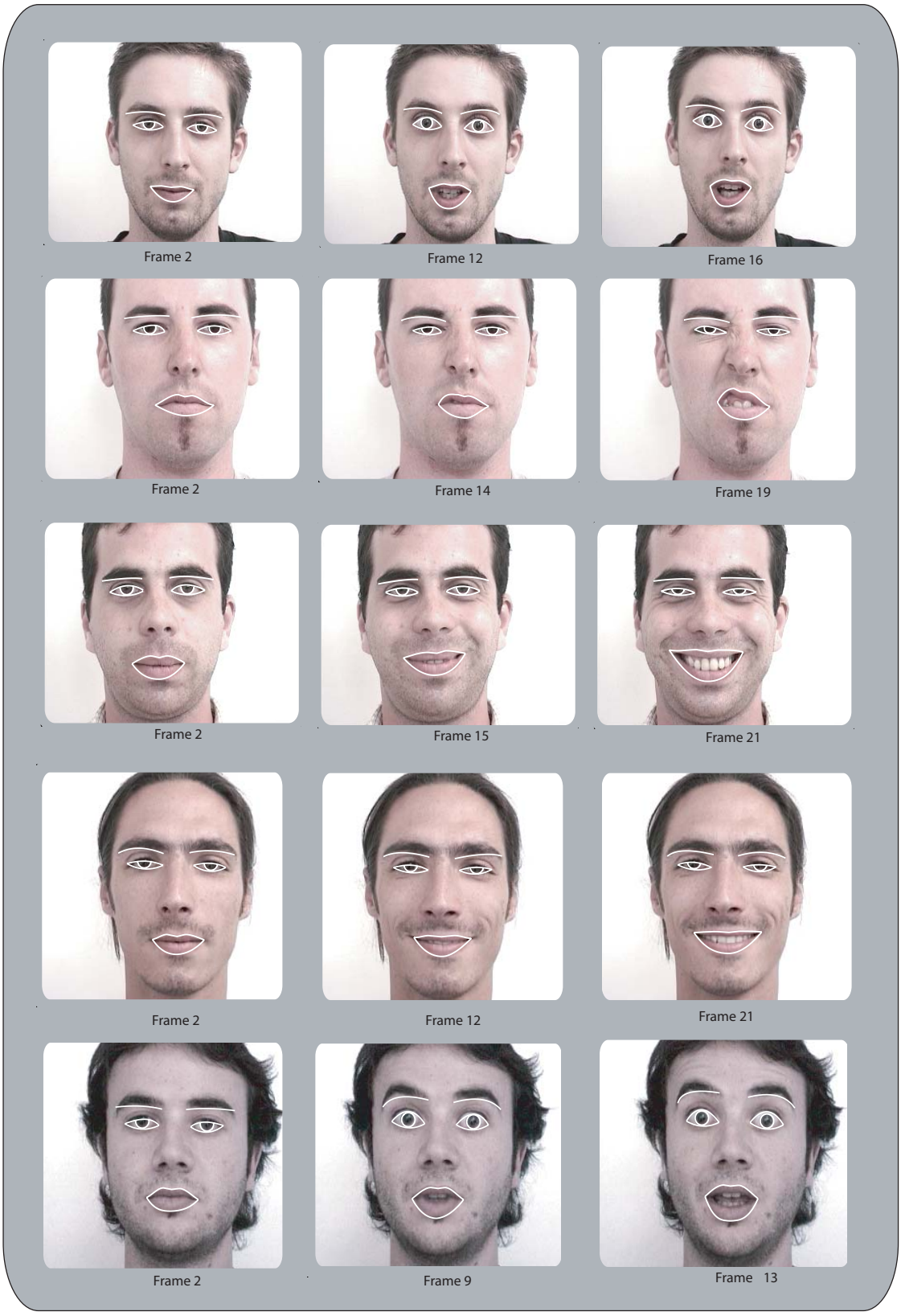

Figure 2.71: Eyes, eyebrows and lips segmentation results.

Part II

Facial expressions classification based on the Transferable Belief Model

Préambule. L'analyse des expressions faciales peut être utile dans plusieurs applications dans différents domaines de l'art, des interactions sociales, des sciences cognitives, de la médecine, de la sécurité ou des interactions homme-machine (IHM). En effet, les expressions faciales procurent non seulement des informations sur l'état affectif, mais aussi sur l'activité, le tempérament et la personnalité [Don99].

D'un point de vue scientifique, plusieurs questions sur les expressions faciales restent sans réponses et quelques pistes sont relativement inexplorées. Par exemple déterminer quelles sont les caractéristiques utilisées par le système visuel humain pour reconnaître les expressions faciales et comment varient ces informations en fonction de l'expression restent des questions ouvertes [Smi05]. Une expression permet de révéler les caractéristiques d'une personne, un message sur l'état interne de la personne. Dans le contexte d'une communication non verbale, une expression implique un changement dans le temps de certains indices visuels dans le visage. Mais de la même manière que l'impression produite par une peinture, les caractéristiques statiques du visage peuvent aussi donner une information sur l'expression.

D'un point de vue vision par ordinateur, la mise en place d'un système automatique de reconnaissance d'expressions faciales est un challenge du fait des différentes contraintes induites par l'application dans un contexte réel. Un tel système doit être d'une grande précision et d'une grande robustesse sans pour autant imposer beaucoup de contraintes à l'utilisateur.

Dans la section qui suit nous décrivons le contexte de la recherche sur la reconnaissance des émotions humaines et nous présentons les expressions comme un indice valide pour les deviner. Nous présentons aussi comment ce problème a été traité en sciences de l'ingénieur et les différents systèmes existants. Ensuite nous présentons l'approche que nous avons mise en place pour reconnaître les expressions faciales. L'approche proposée est basée sur l'analyse des déformations des traits du visage. Le Modèle de Croyance Transférable est utilisé comme processus de fusion.

Preamble. The study of human facial expressions has many impacts in several domains such as art, social interaction, cognitive science, medicine, security or human-computer interaction (HCI). Indeed facial expressions provide information not only about affective state, but also about cognitive activity, temperament and personality, truthfulness and also psychopathological patterns [Don99].

As a scientific problem, many questions about facial expressions remain unanswered and some areas are relatively unexplored. For example determining which face characteristics the human visual system makes use to get the impression of an expressive face and how this information varies with the expression remains an open question [Smi05]. Expression implies a revelation about the characteristics of a person, a message about something internal to the expresser. In the context of a nonverbal communication, expression usually implies a change of some visual patterns in the face over time. But similarly to the impression produced by a static painting of a mood expression or the capture of a feeling, static characteristics of the face can also provide information about the expression.

From a computer science point of view, developing automatic facial expressions recognition systems is a challenging issue due to the many constraints introduced by the application in a real context. Such systems should provide a great accuracy and robustness without imposing much constraints to the user.

In the following sections we present the context of the research on human emotion recognition and present facial expressions as a valid cue for guessing it. We also present how this problem has been tackled in computer science and the different existing systems. Then we present the system we have developed to recognize facial expressions. The proposed system is based on the analysis of facial features deformations. The Transferable Belief Model (TBM) are used as a powerful fusion process.

Chapter 3

Facial expressions classification based on static data

3.1 Emotion versus facial expression

Emotion is one of the most controversial topics in psychology, a source of intense discussion and disagreement from the earliest philosophers and other thinkers to the present day. Emotion can be described relatively to different components on the basis of physiological or psychological factors, including emotion elicitors, emotion neural processes and emotion faces. There is a long history of interest in the problem of recognizing emotion from facial expression in several disciplines: philosophy (René Descartes), biology (Charles Darwin), and psychology (William James, Paul Ekman).

Since 1649 Descartes [Des49] introduced the six "simple passions": *Wonder*; *Love*; *Hatred*; *Desire*; *Smile*; and *Sadness* and assumed that all the others are composed of some of these six. In 1872 Darwin [Dar72] argued that there are specific inborn emotions and that each emotion includes a specific pattern of activation of the facial expression and behavior. Inspired from the works of Darwin, Ekman, Friesen and Ellsworth [Ekm72], [Ekm82] showed that observers could agree on how to label both posed and spontaneous facial expressions in terms of either emotional categories or emotional dimensions across cultures. Ekman showed pictures of facial expressions to people in the U.S., Japan, Argentina, Chile and Brazil and found that they judged the expressions in the same way. Similar facial expressions tend to occur in response to particular emotion eliciting events. But this was not conclusive because all these people could have learned the meaning of expressions by watching TV. Then the experiment need visually isolated people unexposed to the modern world and the media. Ekman found them in the highlands of Papua New Guinea. The experiment results show that subjects judge the proposed expressions in the same way, moreover, their response to a particular emotion corresponds to the same expression. Based on these results Ekman confirmed the universality of the emotional expressions and put a list of six basic emotional expressions namely *Surprise*, *Anger*, *Disgust*, *Happiness*, *Sadness* and *Fear* (see Figure 3.1).
However, the term "*expression*" implies the existence of something that is expressed and people can cheat about their internal feeling by simulating another expression (like actors). So there is a difference between *facial expression* which can be recognized only by the analysis of the facial features and "*emotion*" which corresponds to an internal feeling and requires more than

features deformations to be recognized. Then the recognition of one facial expression allows to obtain some information about the emotional state but is not sufficient to confirm it and may require other modalities (for example the voice, the gesture) to be recovered.

Figure 3.1: The six universal emotional expressions in the following order: *Happiness*, *Fear*, *Disgust*, *Anger*, *Sadness* and *Surprise* [Ekm72].

3.2 Facial expression representation

Facial expressions represent an important channel of nonverbal communication. Even though human being has acquired the powerful capacity of a verbal language, the role of facial expressions in interpersonal interactions remains substantial and improving these skills is often sought. They are often the basis of significant impressions such as friendliness, trustworthiness or status.

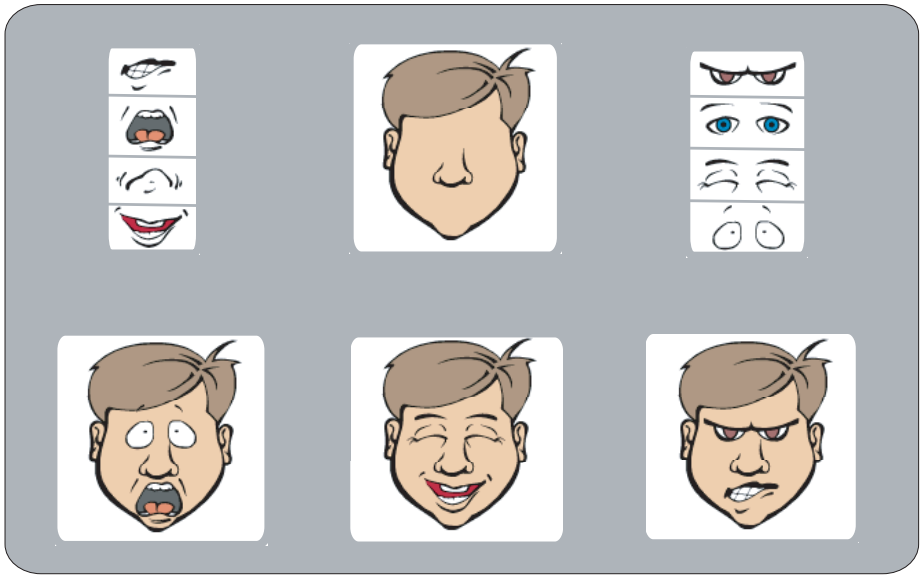

Figure 3.2: Top: information needed for facial expression interpretation; bottom: three examples of facial features (eyes and mouth) configuration leading to *Fear*, *Smile* and *Anger* respectively [Mac06].

The expression of a given face at a specific time is conveyed by a composite of signals from several sources of facial appearance. These sources include the general shape, orientation (pose), and position of the head, the shapes and positions of facial features (e.g., eyes, mouth) and the presence of wrinkles and their shapes. Above all, the most important source is the facial features behavior. For example Figure 3.2 top shows that the face without the permanent facial features does not convey any expression; while each specific combination of eyes and mouth shapes leads to a specific facial expression (see Figure 3.2 bottom).

However, the changes in the facial features appearance are the result of muscular movements produced by part of facial muscles. The facial muscles are like elastic sheets that are stretched in layers over the cranium, facial bones, the openings they form, the cartilage, fat and other tissues of the head. Figure 3.3 shows a 3/4 view of the facial muscles. Each facial expression corresponds to a combination of these facial muscles.

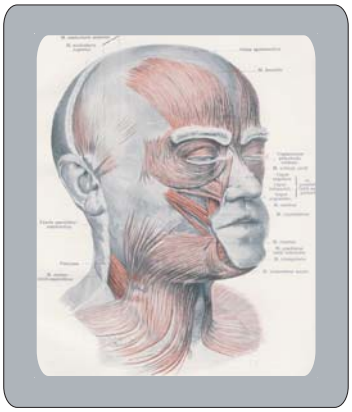

Figure 3.3: Facial muscles involved to produce facial expressions.

In order to code all the possible changes of face appearance, Ekman and Friesen developed the Facial Action Coding System (FACS) in the 1970s. This system became the most widely used and versatile method for measuring and describing facial deformations. The FACS defines the specific changes that occurred with muscular contractions and how best to differentiate one from another. They associated the appearance changes with the action of muscles by the study of the human face anatomy. The FACS measurement units are Action Units (AUs). Ekman defined 46 AUs which represent a contraction or relaxation of one or more muscles (in the case where the appearance change could not be distinguished from the action of the different muscles). Figure 3.4 shows some example of AUs. The FACS coder consists in describing each facial expression as a combination of one or more specific AUs.

Adding to the FACS there exist other coding model, such as MPEG4 standard [Tek99]. MPEG4 is an object-based multimedia compression standard, which allows for encoding of different audiovisual objects in the scene independently. It specifies a face model in its *Neutral* state with a set of feature points: the Facial Definition Parameter set (FDPs) (Figure 3.5 (c) and (d)). The main purpose of these feature points is to provide the Facial Animation Parameter set (FAPs). FAPs represent a set of basic facial actions (tongue, eye and eyebrows motion) and allow the representation of facial expressions. Translational motion parameters are expressed in terms of the Facial Animation Parameter Units (FAPUs). The FAPUs are illustrated in Figure 3.5.a and correspond to fractions of distances between some key facial

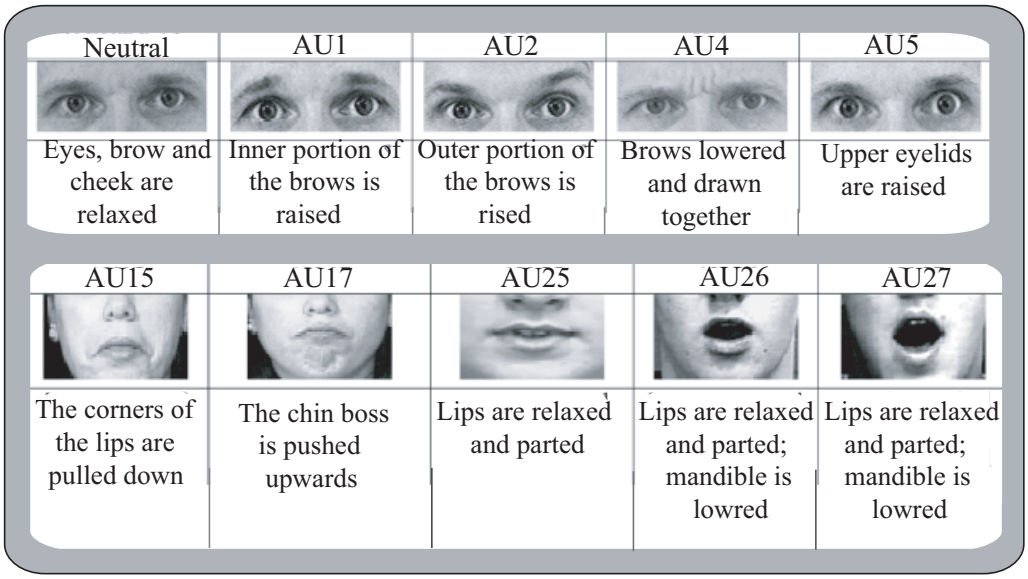

Figure 3.4: Examples of Action Units (AUs) defined by the FACS system. Fist row: upper AUs, second row: lower AUs [Ekm78].

features. These units are defined according to the distances in a *Neutral* expression in order to allow interpretation of the FAPs. Then the use of the MPEG4 standard consists in defining facial expressions as a set of measurements (FDPs) and transformations (FAPs).

To summarize, different coding of the facial feature behaviors can be chosen depending on their suitability relatively to the context of the application. So, an automatic analysis of the emotional expression of a human face requires a number of preprocessing steps to extract information from the face. This information corresponds to the detection and tracking of the face; the localization of a set of facial features (such as eyes, mouth, nose and wrinkles); their representation (for example FACS or MEG4); their interpretation and finally the expression recognition.

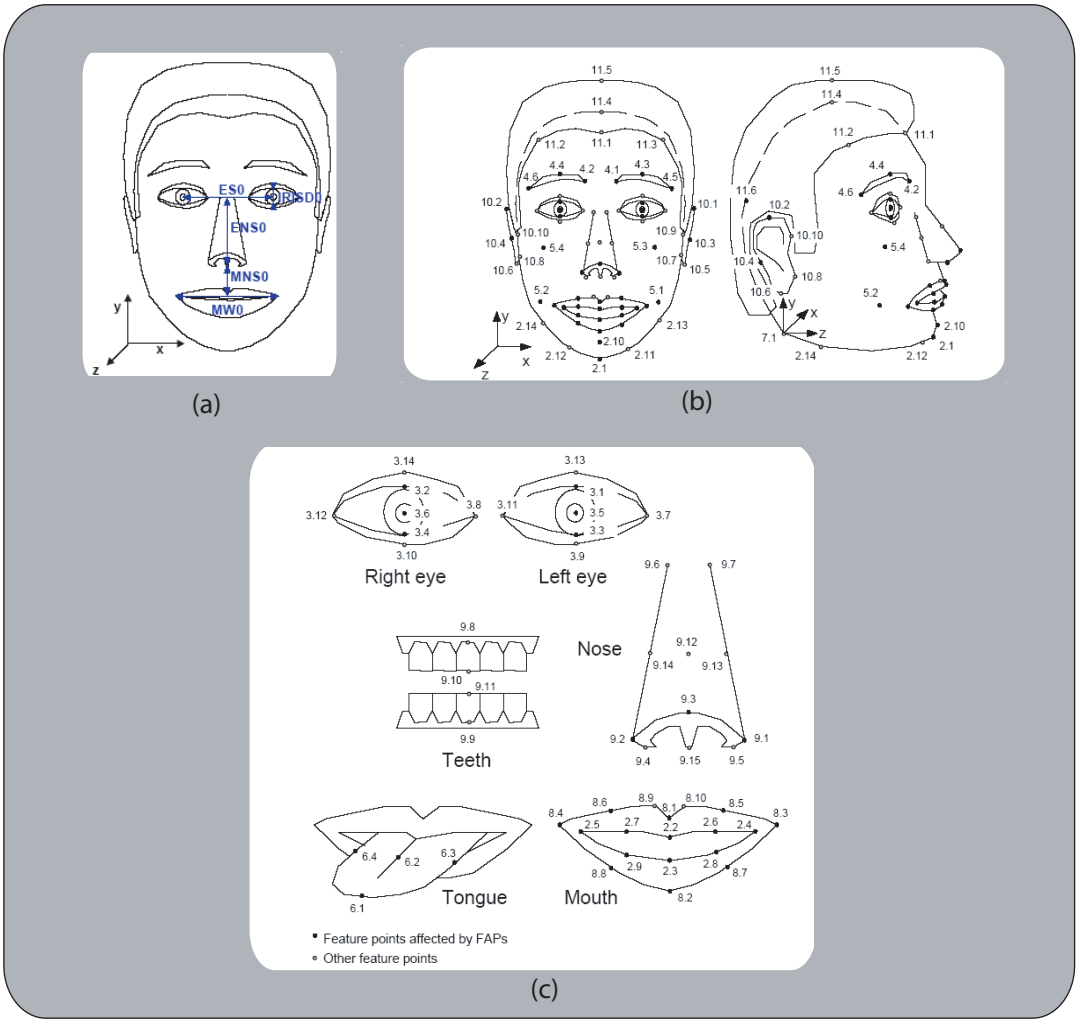

Figure 3.5: (a) A face model in its *Neutral* state and the Facial Animation Parameter Units ; (b) and (c) Facial Definition Parameters used for the definition of the Facial Animation Parameters [Tek99].

3.3 Facial expression recognition based on computer vision

In computer vision, significant amount of research on facial expressions classification have led to many systems adopting different approaches. A detailed description can be found in Pantic et al. [Pan00b] and Fasel et al. [Fas03]. It exists three main approaches: optical flow analysis from facial actions, model based techniques and fiducial based methods.

In the following we present the principles of these methods and we conclude by a comparison table of their results.

3.3.1 Optical flow based approaches

Precise motion information can be obtained by computing optical flow, which represents the direction and magnitude of motion. Several attempts to recognize facial expressions have

focused on optical flow analysis from facial action where optical flow is used to either model muscle activities or estimate the displacements of feature points.

Yacoob and Davis [Yac96] proposed a representation of the facial motion based on the optic flow to recognize the six universal facial expressions. This approach is divided into three stages: first, rectangular regions enclosing the permanent facial features (eyes, eyebrows, nose and mouth) are assumed to be given in the first frame and are tracked in the remaining frames of the sequence (see Figure 3.6); secondly an optical flow estimation on these features define the mid-level representation which describes the observed facial changes at each frame according to the first frame (rigid and nonrigid motion); thirdly this mid-level representation is classified into one of the six facial expressions using a rules based system that combines basic actions of feature components [Bas78] and the rules of motion cues as described in [Ekm78].

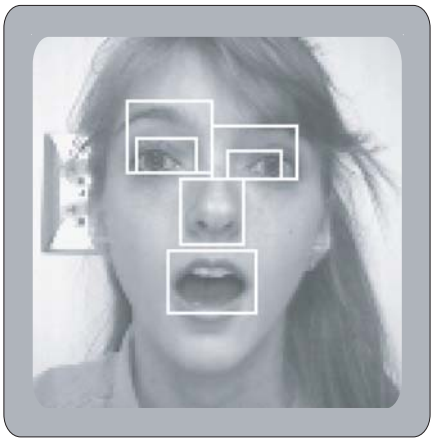

Figure 3.6: An example of rectangles surrounding the face regions of interest[Yac96].

Rosenblum *et al.* [Ros96] expanded the above system with a radial basis function neural network for each one of the six universal facial expressions, which allows to learn the correlations between facial motion patterns and facial expressions. However, the performances of the proposed classification are tested only for *Smile* and *Surprise* expressions. To improve the precision of the model and in order to be robust to head motion, Black and Yacoob [Bla97] have presented an approach with local parameterized model of image motion for facial expression analysis. Rigid head motions are represented by a planar model to recover the information on head motion (see Figure 3.7). Facial features motion is determined relatively to the face. Non-rigid motions of the facial features (eyes, eyebrows and mouth) are represented by an affine plus curvature model. A set of parameters estimated from the models by a regression scheme [Bla93] based on the brightness constancy assumption is used to define mid-level predicates which describe the facial features motion. A set of rules is then defined to combine the mid-level predicates between the beginning and the end of the expressions to recognize it (detailed description of the rules is in [Bla97]). In this approach the initial regions for the head and the facial features are manually selected and are automatically tracked in the remaining frames of the sequence. Moreover the thresholds used to detect the mid-level predicates depend on the face size.

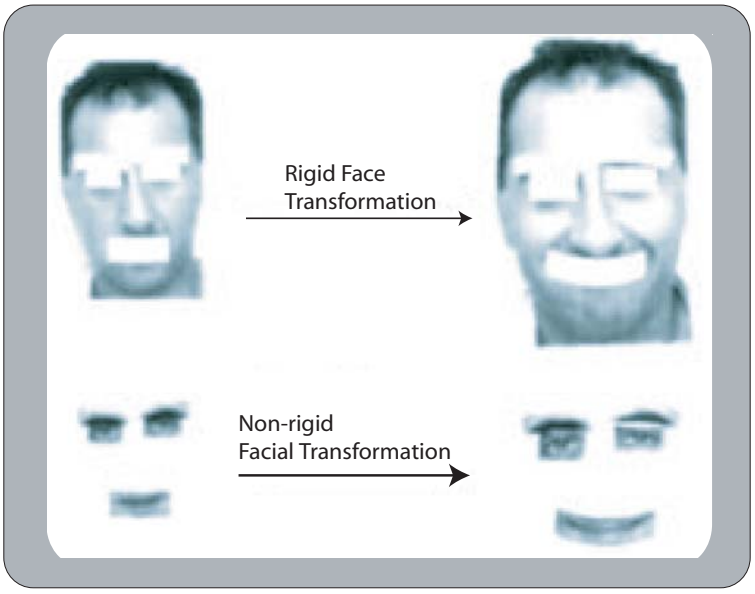

Figure 3.7: Planar model for representing rigid face motions and affine-plus-curvature model for representing nonrigid facial motions [Bla97].

Essa and Pentland [Ess97] have proposed the combination of a dynamic physical model and of motion energy for facial expressions classification. Motion is estimated from optical flow and is refined by the physical model in a recursive estimation and a control framework. A physical face model is applied to model facial muscle actuation and an ideal 2D motion is computed for the five studied expressions (*Anger*, *Disgust*, *Happiness*, *Surprise* and *raised eyebrows* (the other expressions are difficult to simulate by their subjects) (see Figure 3.8). Each template has been delimited by averaging the patterns of motion generated by two subjects for each expression. Facial expressions classification is based on the Euclidean distance between the learned template of motion energy and the motion energy of the observed image.

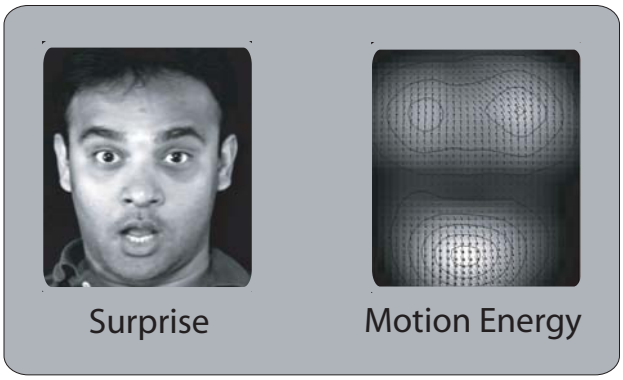

Figure 3.8: The spatio-temporal motion energy representation of facial motion for surprise [Ess97].

Since their approach needs a physical face model, the classification accuracy relies on the validity of such a model which appeared difficult to design.

Cohn *et al* [Coh98] propose an automatic recognition system based on the AUs modeling. The displacement of 36 manually located feature points around eyes, eyebrows, nose and mouth (see Figure 3.9) are estimated using optical flow. Separate group (variance-covariance matrices) were used for the classification of the AUs. They used two discriminant functions for three AUs of the eyebrow region, two discriminant functions for three AUs of the eye region and five discriminant functions for nine AUs of the nose and mouth regions. However, optical flow is computed at each pixel within a specified region of interest. This approach does not always allow to distinguish between the optical flow caused by facial features motion and the one caused by other unrelated noise, leading to false detection results. For example, in the case of *Surprise*, the detection of a mouth area corresponding to an *open* state can be due to the error on the estimation of the optical flow caused by the luminance variation. Moreover, optical flow estimations are easily disturbed by nonrigid motion. They are also sensitive to the inaccuracy of the image registration and to motion discontinuities.

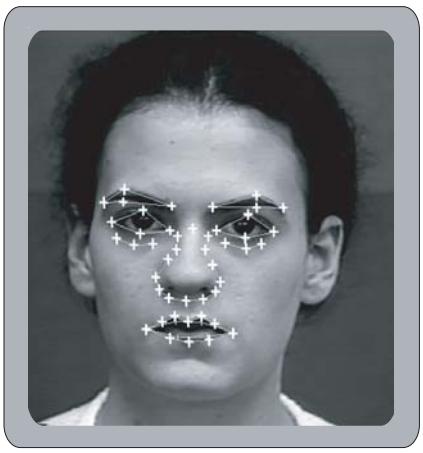

Figure 3.9: Feature point tracking [Coh98].

3.3.2 Model based approaches

Several facial expression recognition systems have employed model-based techniques. Some of them apply an image warping process to map face images onto geometrical model. Others realize a local analysis where spatially localized kernels are employed to filter the extract facial features. A number of works applied holistic gray level analysis based on principal components analysis, Gabor wavelet analysis, or Eigenface and Fisherface approach.

Huang and Huang [Hua97] first compute 10 Action Parameters (APs) (see Figure 3.10) based on the difference between the model feature parameters in an neutral face and those in the examined facial expression of the same person. The first two terms of eigenvalues are used to represent the APs variations. Then a minimum distance classifier was used to cluster the two principal action parameters of 90 training image samples into six clusters (the six basic emotional expressions). Since the principal component distribution of each expression is overlapped with the distribution of at least two other expressions, three best matches are

selected. The highest score of the three correlations determines the final classification of the examined expression. However, the tests are achieved on the same subject then it is not known how the method will behave for an unknown subject.

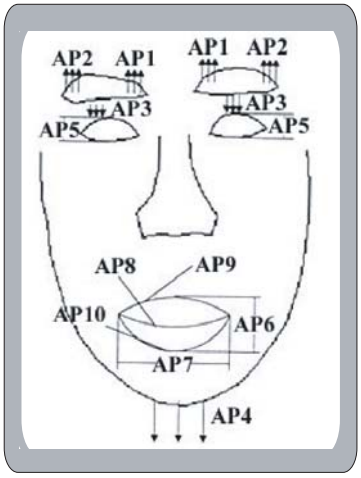

Figure 3.10: APs [Hua97].

Based on the Gabor wavelet representation, Lyons [Lyo98] presented a method for the classification of the six universal expressions plus the *Neutral* expression. A grid of 34 facial feature points is manually initialized on the face (Figure 3.11). The Gabor wavelet coefficients of each feature point of the grid are computed and combined into a single vector. The principle components of the feature vectors from training images are computed. Then linear discriminant analysis is used in order to aggregate the resulting vectors into clusters having different facial attributes. Finally, classification was performed by projecting the input vector of a test image along the discriminant vectors. The input vector is affected to the nearest cluster.

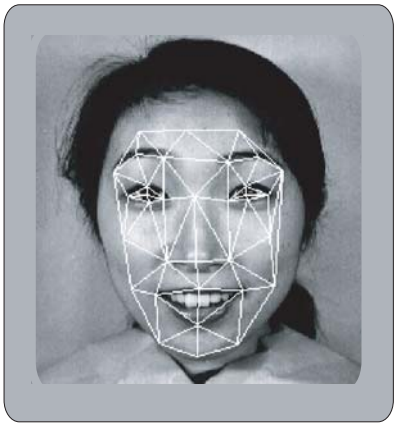

Figure 3.11: The Gabor-labeled elastic graph representation of facial image [Lyo98].

Based on the work of Lyons, Zhang [Zha98] classifies the same expressions using a neural network. The inputs of the network correspond to the positions of the 34 facial feature points defined in [Lyo98] with the response of 18 Gabor wavelets. In the classification step each output unit gives an estimation of the probability of the examined expression belonging to the associated category.

Oliver *et al* [Oli00] applied a Hidden Markov model (HMM) on facial expression recognition based on the deformation of mouth shapes tracked in real-time. The mouth shape is characterized by its area, its spatial eigenvalues (e.g., width and height) and its bounding box. Figure.3.12 depicts the extracted mouth feature vector. Based only on the mouth shape the studied expressions are *Open mouth*, *Sadness*, *Smile* and *Smile-Open mouth*. Each one of the mouth-based expressions is associated to a HMM trained on the mouth features vector. The facial expression is identified by computing the maximum likelihood of the input sequence with respect to all trained HMMs. However, only a part of the facial expressions have the characteristic pattern contained in the mouth shape.

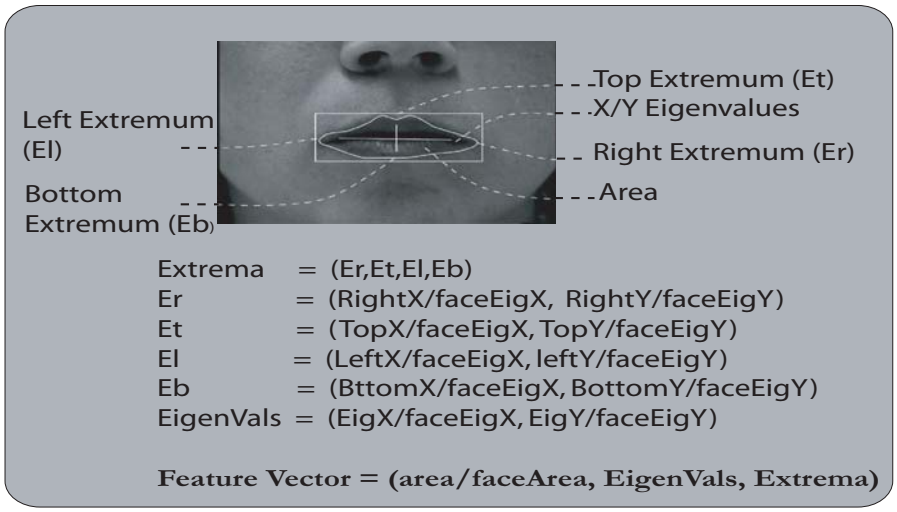

Figure 3.12: Mouth feature vector extraction [Oli00].

Dubuisson [Dub02] first performs a dimensionality reduction by applying PCA to the face. Then the method searches for the most discriminant projection along eigenvectors by sorting the principal components after PCA and keeping the K most discriminant in their importance order for the recognition task. Finally, LDA is computed into this sorted eigenspace, to generate a Fisherspace where new samples are classified. Figure 3.13 shows the five generated Fisherfaces applied to a learning set containing six facial expression classes. Then, a decision tree classifier [Cho91] is used as a measure between the projection of the tested samples and each estimated mean vector in the Fisherspace. The performances of the recognition method are strongly dependent on the precision of the postprocessing step where the normalization and registration of each new face are required and manually achieved.

Gao [Gao03]proposes classification approach of three expressions *Neutral*, *Smile* and *Scream* applying line-based caricatures. Their approach uses line edge map (LEM) [Gao99] as expression descriptor (see Figure 3.14.a). The classification is obtained by computing Haus-

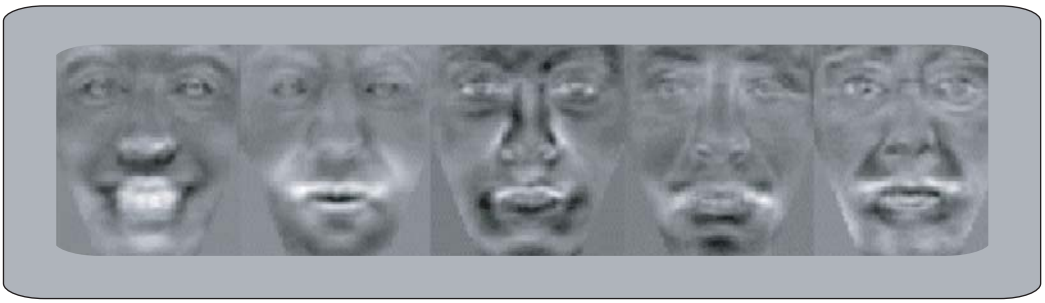

Figure 3.13: Illustration of the five Fisherfaces, corresponding to five axes of the subspace generated by sorted PCA plus LDA method, which are used as basis of the final discriminant subspace [Dub02].

dorff distance on the directed line segments (disparity measure defined between two line sets) between the current face LEM and the caricature models of expressions (see Figure 3.14.b).

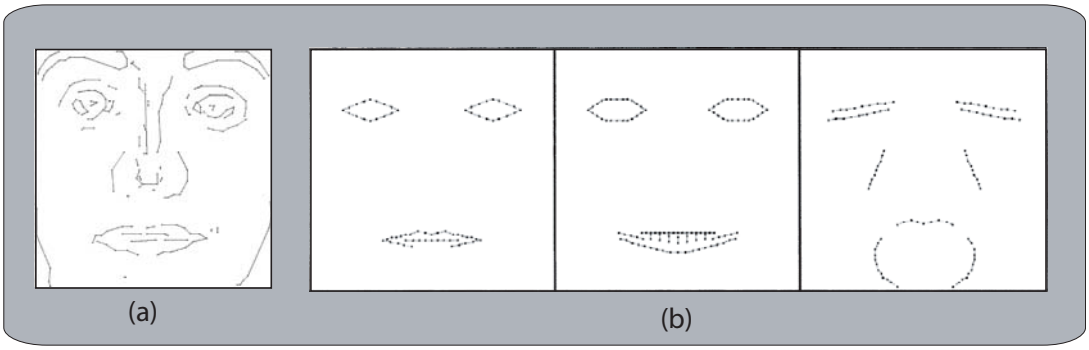

Figure 3.14: (a) LEM of a face; (b) facial expression models [Gao03].

Based on the active appearance model, Abboud *et al* [Abb04b] proposed a model for automatic facial expressions classification. Each face image is represented by a corresponding appearance vector [Abb04a]. Then, Mahalanobis distance [Han81] is measured between the tested appearance vector and each estimated mean vector in Fisherspace. In each configuration, the tested face is assigned to the class having the nearest mean.

Anderson [And06] proposed an automated multistage system for real-time recognition of facial expression. It uses facial motion to characterize monochrome frontal views of facial expressions and is able to operate effectively in cluttered and dynamic scenes. Faces are located using a spatial ratio template tracker algorithm. Optical flow of the face is subsequently determined using a real-time implementation of a gradient model. The expression recognition system then averages facial velocity information over identified regions of the face and cancels out rigid head motion by taking ratios of this averaged motion. The motion signatures produced are then classified using Support Vector Machines as either non expressive or as one of the six universal expressions. However, the system is specific to a single user at a fairly well fixed distance and aspect to the camera.

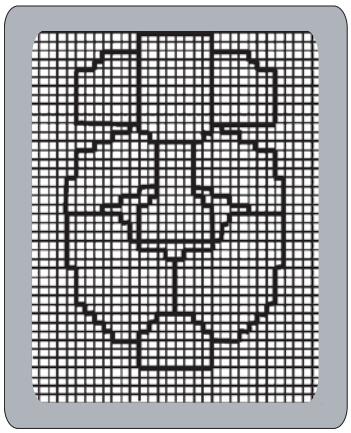

Figure 3.15: Regions for motion averaging [And06].

It is difficult to design a deterministic physical model that accurately represents facial geometrical properties and muscle activities. The holistic approach usually involves a time intensive training stage. The trained model is often unreliable for practical uses due to interpersonal variations, the fact that expressions are acted, illumination variability and also difficulties to cope with dynamic emotional sequences.

3.3.3 Fiducial points based approaches

Recent years have seen the increasing use of geometrical features analysis to represent facial information. In these approaches, facial movements are quantified by measuring the geometrical displacements of facial feature points between the current frame and the initial frame.

Lien *et al* [Lie98] propose an hybrid method based on: first, the feature points tracking [Luc81] (points around the contours of the eyes, eyebrows, nose and mouth manually detected in the first frame), second, optical flow and third, furrow detection to extract expression information. Expression classification is based on Facial Action Coding System (FACS) action units (AUs) [Ekm78]. HMMs are used for the discrimination between each AUs or AUs combinations according to the pattern of features motions. A directed link between states of the HMM represents the possible inherent transition from one facial state to another. An AU is identified if its associated HMM has the highest probability among all HMMs given a facial feature vector. The main drawback of the method is the number of HMMs required to detect a great number of AUs or combination of AUs involved in the facial expressions classification.

Tian *et al* [Tia01] use two separate Neural Networks (NNs) based approach to recognize 6 upper and 10 lower AUs based on both permanent facial features (eyes, eyebrows and mouth) and transient facial features (deepening of facial furrows). The facial features have been grouped into separate collections of feature parameters because the facial actions in the upper and lower face are relatively independent for AU recognition [Ekm78]. The inputs of the NNs for both training and classification are the parametric descriptions of permanent and transient facial features (see Figure 3.17). The facial features are manually initialized in the first frame and tracked in the remaining frames of the sequence. The facial expression recognition is realized by the combination of the upper face and the lower face AUs.

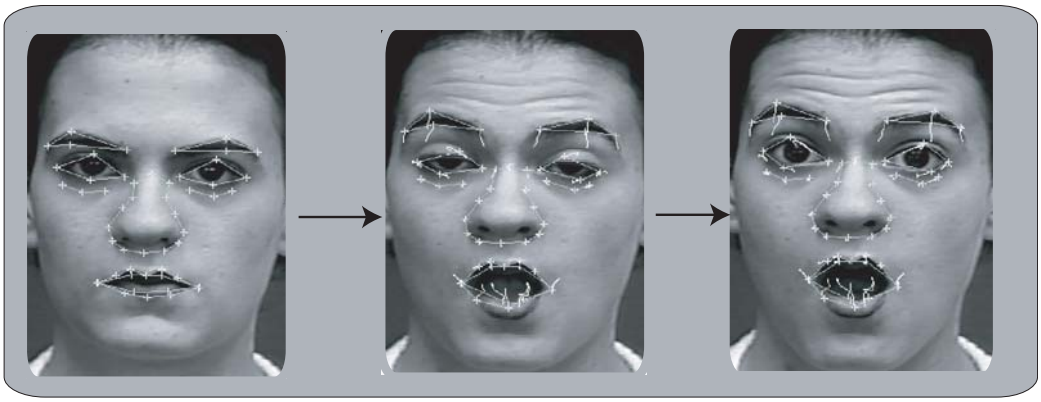

Figure 3.16: Facial feature point tracking sequence [Lie98].

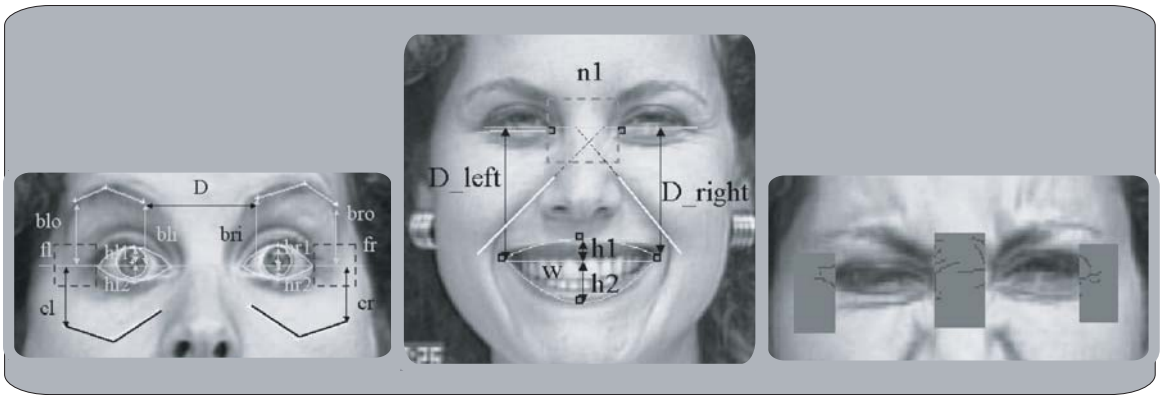

Figure 3.17: Left: Upper face features *hl=(hl1+hl2)* and *hr=(hr1+hr2)* are the height of left eye and right eye; D is the distance between brows; *cl* and *cr* are the motion of the left cheek and right cheek. *bli* and *bri* are the motion of the inner part of left brow and right brow. *blo* and *bro* are the motion of the outer part of left brow and right brow. *fl* and *fr* are the left and right crow's-feet wrinkle areas. Middle: lower face features. *h1* and *h2* are the top and bottom lip heights. w is the lip width. D_{left} is the distance between the left lip corner and eye inner corners line. D_{right} is the distance between the right lip corner and eye inner corners line. *n1* is the nasal root area. Right: Nasal root and crow's-feet wrinkle detection [Tia01].

Pantic and Rothkrantz [Pan00a] use face models made of dual-view points for facial expressions classification: the frontal view and the side view (see Figure 3.18). After automatic segmentation of facial features (eyes, eyebrows and mouth), they code several characteristic points (such as eyes corners, mouth corners,etc) into AUs using a set of rules. Then the FACS [Ekm78] is used to recognize the six universal facial expressions. The classification is performed by comparing the AU-coded description of facial expressions of observed expression against the rule descriptors FACS.

Pardas *et al* [Par02] and Tsapatsoulis *et al* [Tsa00] propose a description of the six universal facial expressions using the MPEG-4 Facial Definition Parameter Set (FDP) [Tek99]

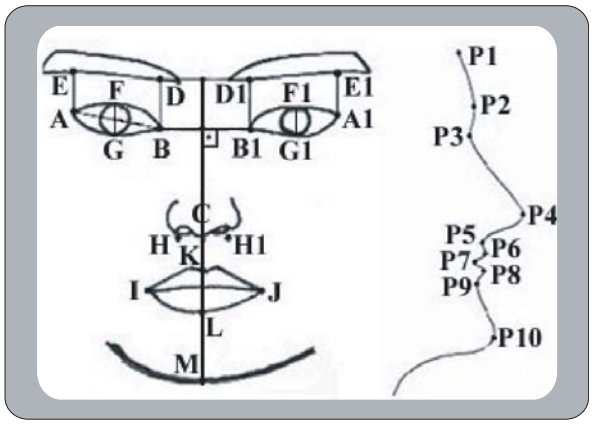

Figure 3.18: Facial points of the frontal-view face model and the side-view face model [Pan00a].

(see Figure 3.5.b). Tsapatsoulis *et al* [Tsa00] use all the FAPs (defined in [Tek99]) and propose a classification based on a fuzzy inference system. Based only on the eyebrows and mouth segmentation contours, Pardas *et al* [Par02] used the corresponding FAPs (8 for the eyebrows and 10 for the mouth) for the classification process. This one is based on an HMM based system which assigns to the input the expression with the highest probability.

Cohen *et al* [Coh03a] developed a system based on a non-rigid face tracking algorithm [Tao98] to extract local motion features (see Figure 3.19). These motion features are the inputs of a Bayesian network classifier used to recognize the six universal facial expressions. The feature-based representation requires accurate and reliable facial feature detection and

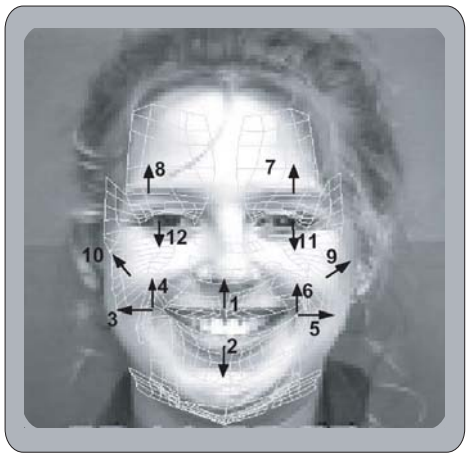

Figure 3.19: The facial motion measurements [Coh03a].

tracking to cope with variation of illumination, significant head movement and rotation, as well as nonrigid feature change.

It is not possible to directly compare the surveyed facial expressions classification systems due to the varying facial action labeling and the different test beds that were used. Table 5.2

summarizes the methods for facial expressions classification, the number of subjects used to make the evaluation and their performances.

3.3.4 Conclusion

The described approaches are similar in a general sense: they first extract some features from the face image; these features are then used as input of a classification system which classifies them as one of the preselected facial expressions. They differ mainly in the way the low level analysis is performed (the features extraction) and in the way the classification is implemented. Tables 3.1, 3.2, 3.3 summarize these facial expressions classification approaches.

In general, the performances are comparable to trained human recognition which achieves about 87% as reported by Bassili [Bas78]. However, they present some limitations in our view. First, most of them require important manual intervention for the detection and accurate normalization of test faces, during the initialization of facial feature tracking approaches. Second, most of them map facial expressions directly into one of the basic facial expression proposed by Ekman and Friesen [Ekm78]. This leads to two main problems. On the one hand, since people are not binary, pure facial expressions are rarely produced. On the contrary they rather move from one expression to another, transiting by mixed expressions. Therefore, the classification of an expression into a single emotion category is not realistic and, ideally, the classification system should be able to identify such intermediate mixture of expressions. On the other hand, the classification into one the 6 basic facial expressions is not always realistic because a human face can display numerous facial deformations which may not correspond to any of these expressions. Then the system should be able to take into account these unknown expressions.

For all these reasons, we propose a facial expressions classification system based on the use of the Transferable Belief Model (TBM) [Sme98].

Reference	Classification	Nbr of AUs or expressions	Nbr of test subjects	Performances (%)
Optical flow based approaches				
[Yac96]	rule-based	6	32 subjects: sm(37), an(24) su(30), di(13) fe(7), sa(5)	88
[Ros96]	neural networks	2	32	88
[Bla97]	rule-based	6	40 subjects: su(35), an(20) sa(8), sm(61) fe(6), di(15)	88
[Ess97]	spatio-temporal motion-energy templates	5	7 subjects: su(10), rai_eye(10) di(10), sm(12) an(10)	98
[Coh00]	Discriminant functions	8 AUs + 7 AUs combination	100 subjects Cohn_Kanade database	88

Table 3.1: Comparisons of facial expression recognition algorithms adopting on optical flow based approaches. sm: *Smile*, an: *Anger*, su: *Surprise*, di: *Disgust*, fe: *Fear*, sa: *Sadness*, AUs: Action Units.

Reference	Classification	Nbr of AUs or expressions	Nbr of test subjects	Performances (%)
Model based approaches				
[Hua97]	2D emotion space (PCA) minimum distance classifier	6	1 subject	84.5
[Lyo98]	PCA and LDA of the labeled graph vectors	7	9 Japanese females	75 – 92
[Zha98]	neural networks	6	9 Japanese females	90.1
[Oli00]	HMM	4	8 subjects	95.95
[Dub02]	PCA and LDA and Tree-based classifier	6	Cohn_Kanade database fe(38), di(28) su(88), an(27) sa(55), sm(109)	87.6
[Gao03]	distance-based	3	112 subjects (61 males + 51 females)	86.6
[Abb04b]	PCA and LDA Mahalanobis distance classifier	7	Cohn_Kanade database: fe(17) sm(26), su(23) sa(18), de(20) ne(45), an(17)	83.73
[And06]	SVM	7	–	81.82

Table 3.2: Comparisons of facial expression recognition algorithms adopting model based approaches. sm: *Smile*, an: *Anger*, su: *Surprise*, di: *Disgust*, fe: *Fear*, sa: *Sadness*, AUs: Action Units.

Reference	Classification	Nbr of AUs or expressions	Nbr of test subjects	Performances (%)
Fiducial points based approaches				
[Lie98]	HMMs	3 Upper AUs + 6 Lower AUs	85 subjects Cohn_Kanade database	Upper: 85 (features points, edge density) 93 (dense flow tracking) Lower: 88 (features points) 81 (edge density)
[Pan00a]	rules based expert system	6	8 subjects	91
[Tsa00]	fuzzy inference	6		81.16
[Tia01]	2 neural networks (Upper and Lower)	6 Upper AUs + 10 Lower AUs and *Neutral* expression	Cohn_Kanade database: 14 Upper 32 Lower	Upper 96.4 Lower 96.7
[Par02]	HMM	6	Cohn_Kanade database:	84
[Coh03a]	bayesian network	6	5	86.45

Table 3.3: Comparisons of facial expression recognition algorithms adopting fiducial points based approaches. sm: *Smile*, an: *Anger*, su: *Surprise*, di: *Disgust*, fe: *Fear*, sa: *Sadness*, AUs: Action Units.

3.4 Our contribution

As explained before, from a physiological perspective, a facial expression results from the deformations of some facial features caused by an emotion [Ekm99]. Each emotion corresponds to a typical stimulation of the face muscles; thereby deformations of facial features like eyes, eyebrows or mouth. These deformations form all together the so-called facial expression.

The aim of our work is to evaluate the possibility of recognizing the six universal emotions by only considering the deformations of permanent facial features such as eyes, eyebrows and mouth.

Figure 3.20 presents skeletons of expression, i.e images of faces displaying only the contours of the eyes, the eyebrows and the mouth.

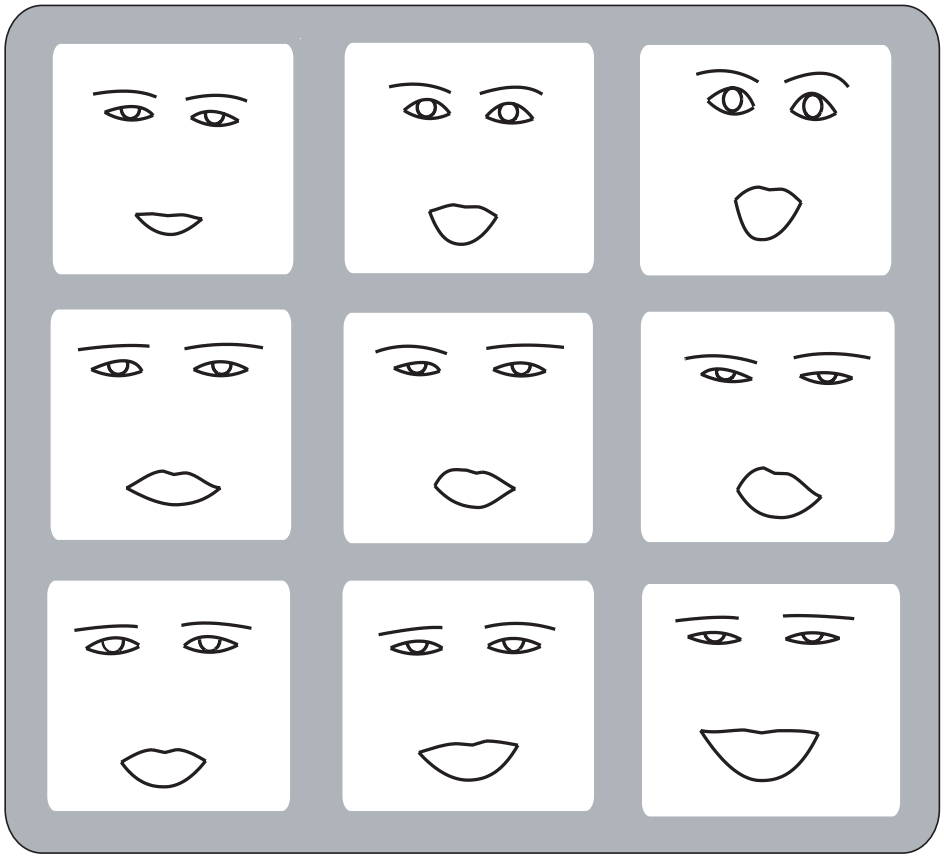

Figure 3.20: Skeletons of expressions: sequence of *Surprise* (top); sequence of *Disgust* (middle); sequence of *Smile* (bottom).

At this point, we make the hypothesis that these features contain enough information to recognize the considered expression.

To validate this hypothesis from a human classification viewpoint, we carried out an experiment in collaboration with the laboratory of social psychology (LPS) of Grenoble [Sou03]. Skeleton images corresponding to contours of permanent facial features (see Figure 3.20) were presented to 60 subjects. Each subject was asked to classify each skeleton to one of the six facial expressions (*Smile*, *Surprise*, *Disgust*, *Anger*, *Fear*, *Sadness* and *Neutral*). We registered

60% of good classification. This rate has to be compared to the value of 87% of good classification we obtained when a human classification is performed on the corresponding original images. This suggests that humans are able to identify facial expressions by viewing only the contours of the facial features.

The originality of our approach also consists in proposing a fusion architecture based on the Transferable Belief Model. This fusion method is well suited for the problem of facial expressions classification: this model facilitates the integration of a priori knowledge and it can deal with uncertain and imprecise data which could be the case with data obtained from video-based segmentation algorithms. In addition it is able to model intrinsic doubt which can occur between facial expressions in the recognition process (see Figure 3.21). It allows the classification of different expressive states like "pure" expression and mixture of expressions. Considering that "binary" or "pure" facial expressions are rarely produced (people show a mixture of facial expressions), the classification of any facial expression into a single emotion category is not realistic. Secondly the proposed method is sensitive to different expressions intensities and allows to determine the *Unknown* expressions corresponding to all facial deformations that can not be categorized into one of the predefined facial expressions.

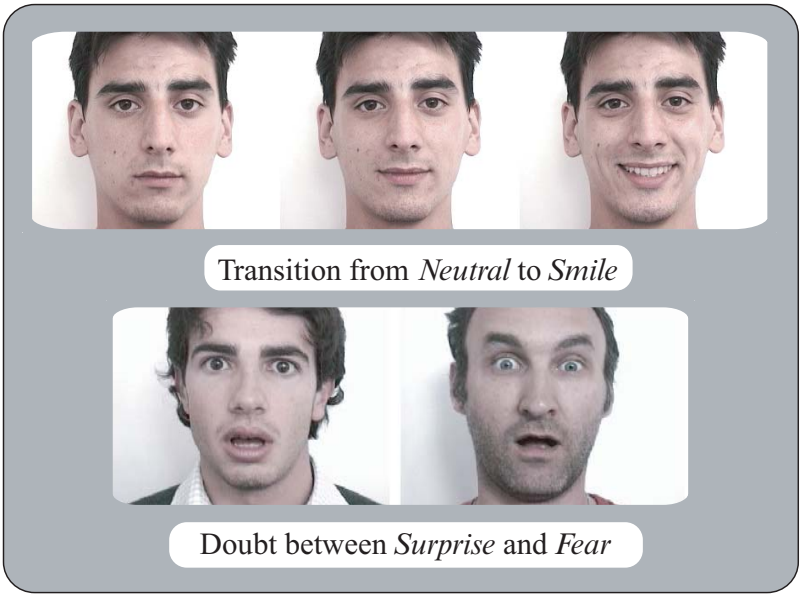

Figure 3.21: Example of doubt between expressions.

3.5 System overview

Our goal is an automatic system for facial expressions recognition in video sequences. In the following we present a facial expressions classification on static images representing frontal viewed faces. The global scheme of the proposed recognition system is described in Figure 3.22. It mainly consists of four processing blocks: segmentation, data extraction, data analysis and classification. In the segmentation step (see Figure 3.22 .a), frontal viewed face images are presented to the system and contours of eyes, eyebrows and mouth are located by using the segmentation algorithm described in part I.

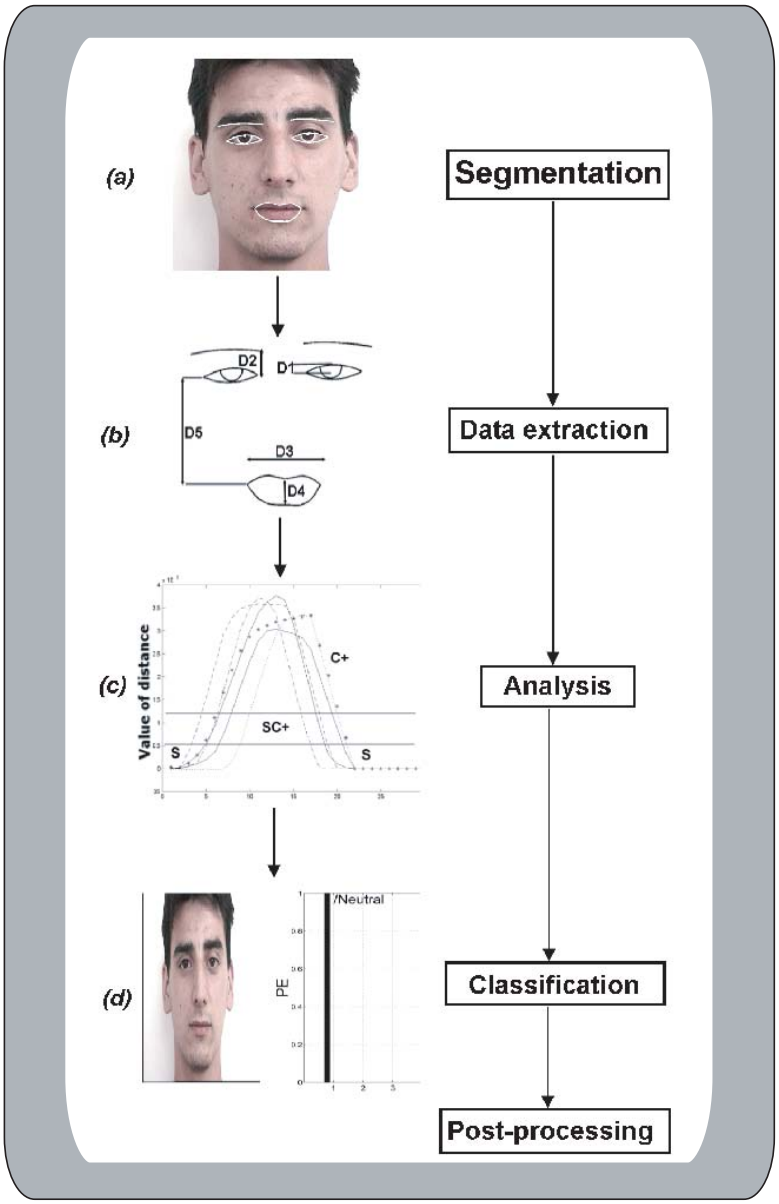

Figure 3.22: Overview of the classification system.

In the data extraction step (see Figure 3.22 .b), a skeleton of the facial features is generated from the contours and all these features are used for the classification process. Several works have been done to define the most pertinent part of the face to recognize each facial expression. Ekman and Boucher [Bou75] conclude that the bottom half part of the face is mainly used by human observers for the recognition of *Smile* and *Disgust* and that the whole face is used for the recognition of *Anger* and *Surprise*. Bassili ([Bas78], [Bas79]) notes that the whole face yields to a better recognition of facial expressions (74.4%) compared to the bottom part of the face only (64.9%) or the top part of the face only (55.1%). Gouta and Miyamoto [Gou00] works conclude that the top half part of the face yields to a better recognition of *Anger*, *Fear*, *Surprise* and *Sadness*, the bottom half part of the face to a better recognition of *Disgust* and

Smile and the whole face, to a better recognition of *Neutral*. Based on these works it seems that the whole face is necessary to recognize all the facial expressions even if in some cases, only a small part of the face could be sufficent.

From the skeleton of the facial features several distances characterizing their deformations are computed. The choice of five facial distances has been motivated by Bassili's work [Bas79]. One challenge is to evaluate the well-founded of these choices. More details can be found in Part II section 3.6.2.

In the data analysis step (see Figure 3.22 .c), the numerical values of the characteristic distances are mapped to symbolic states that qualitatively encode how much a given distance differs from its corresponding value in the *Neutral* state. Then each facial expression is characterized by a combination of characteristic distances states. More details are given in Part II sections 3.7.1 and 3.7.2.

In the classification step (see Figure 3.22 .d), the Transferable Belief Model (TBM) [Sme98] are applied to recognize the facial expressions. It consists in the fusion of the confidence associated to each characteristic distance, in order to find the most believable hypothesis that can be a single facial expression, a mixture of facial expressions or an *Unknown* expression (expression which do not correspond to any of the predefined expressions). More details are given in Part II section 3.7.3.

3.6 Facial data extraction

3.6.1 Facial features segmentation

The extraction of facial feature contours is detailed in Part I but in the following we summarize the main ideas. A specific parametric model is defined for each deformable feature. Several characteristic points are extracted in the image to be processed to initialize each model (for example eyes corners, mouth corners and eyebrows corners). In order to fit the model with the contours to be extracted, a gradient flow (of luminance and/or chrominance) through the estimated contour is maximized because at each point of the searched contour, the corresponding gradient is normal. The definition of a model associated to each feature offers the possibility to introduce a regularisation constraint.

3.6.2 Measures extraction

The segmentation process leads to a skeleton of facial expression (see Figure 3.20). This skeleton is used to determine the facial features deformations occurring when an expression is presented on the face. Five basic characteristic distances named D_1 to D_5 are defined (see Figure 3.23 and Figure 3.24) on each skeleton. These five distances correspond to the mapping of the rules introduced in Bassili's work and to the rule defined in the MPEG-4 description of the deformations undergone by facial features for each expression [Tek99]. These distances are normalized with respect to the distance between the centers of both irises in the analyzed face. This makes the analysis independent on the variability of face dimensions and on the position of the face with respect to the camera.

In addition to distance normalization, only the deformations with respect to the *Neutral* expression are considered. Meaning that each distance D_i is normalized by its corresponding value measured in the *Neutral* expression D_{iN} (N for *Neutral* state).

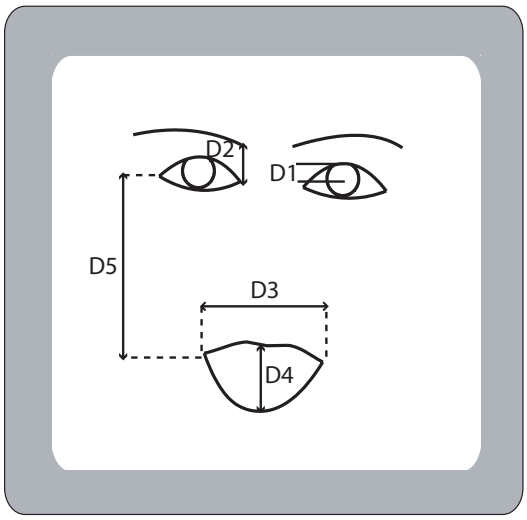

Figure 3.23: Characteristic distances.

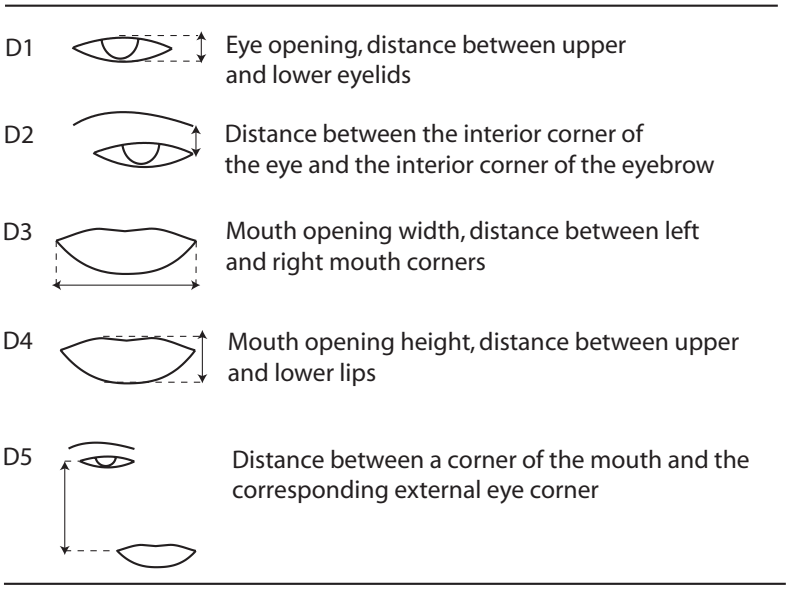

Figure 3.24: Characteristic distances computed on facial skeleton images.

3.7 Facial data classification by the Transferable Belief Model

Our aim is to characterize each facial expression by a specific combination of states associated to the characteristic values of the measured distances D_i. In our application, we have 7 facial expressions: $\Omega_E = \{Smile(E_1), Surprise(E_2), Disgust(E_3), Sadness(E_4), Anger(E_5), Fear(E_6), Neutral\ (E_7)\}$. At any time (or frame), each facial distance should give an idea about the corresponding facial expression. To do this, we propose a two-step procedure: the

first step associates a symbolic state to each distance and the second step defines the rules between the symbolic states and the facial expressions.

3.7.1 Symbolic states

The analysis of the distance values shows that each distance can be higher, lower or roughly equal to its corresponding value in the *Neutral* expression. We associate a state variable V_i(1 $\leq$ i $\leq$ 5) to each characteristic distance D_i in order to convert the numerical value of the distance to a symbolic state. V_i can take 3 possible states { S, C^+, C^- } depending on how different is D_i from its corresponding value in the *Neutral* expression.

- $V_i = C^+$ if the current distance D_i is significantly higher than its corresponding value in the *Neutral* expression;
- $V_i = S$ if the current distance D_i is roughly equal to its corresponding value in the *Neutral* expression;
- $V_i = C^-$ if the current distance D_i is significantly lower than its corresponding value in the *Neutral* expression.

Two undetermined regions corresponding to a doubt between two states are added:

- $V_i = S \cup C^+$, doubt between S and C^+ if the current value of the distance D_i is neither sufficiently high to be C^+ and neither sufficiently stable to be S ($\cup$: logical OR);
- $V_i = S \cup C^-$, doubt between S and C^- if the current value of the distance D_i is neither sufficiently low to be C^- and neither sufficiently stable to be S.

Figure 3.25 shows the temporal evolution of characteristic distances for several video sequences going from *Neutral* to a given expression and coming back to *Neutral*. We observe similar evolutions for the characteristic distances associated to a same facial expression. For example (see Figure 3.25 left), the characteristic distance D_2 (distance between eye corner and eyebrow corner) always increases in case of *Surprise* because people have wide opened eyes, so that the state variable V_2 evolves from the equal state (S) to the significantly higher state (C^+) via an undetermined region ($S \cup C^+$). For the other example (see Figure 3.25 right), the characteristic distance D_5 (distance between mouth corner and eye corner) always decreases in the case of *Smile* because people have opened the mouth and the mouth corners get closer to the eyes. Thus the state variable V_5 goes from the equal state (S) to the significantly lower state (C^-) via an undetermined region ($S \cup C^-$).

3.7.2 Logical rules between symbolic states and facial expressions

Figure 3.26 shows how the characteristic distances are typically mapped to the symbolic states with respect to the facial expressions. This mapping has been obtained by heuristic analysis of the HCE database (see section 1.1.4) for *Smile*, *Surprise*, *Disgust* and *Neutral* expressions. The proposed combinations of symbolic states associated to each D_i for the 4 expressions *Smile*, *Surprise*, *Disgust* and *Neutral* are compared to the MPEG-4 description of the deformations undergone by facial features for such expressions [Tek99]. As a result, we find that the proposed combinations are compliant with MPEG-4 description.

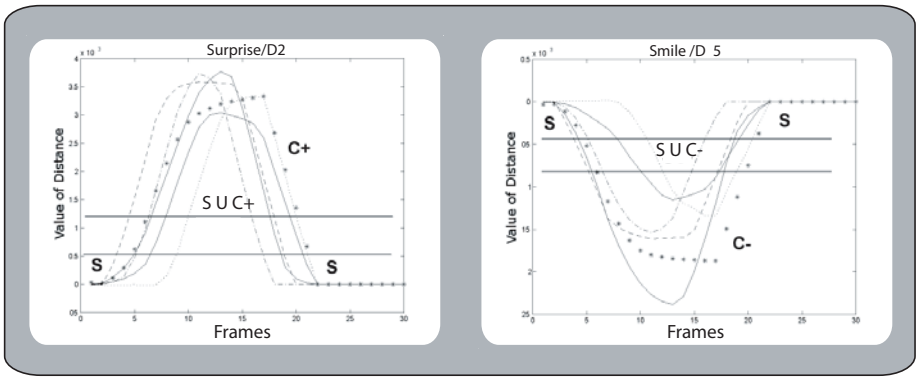

Figure 3.25: Time evolutions of characteristic distances and corresponding state values for: left D_2 in case of *Surprise* and right D_5 in case of *Smile* for several subjects (one curve per subject).

	V1	V2	V3	V4	V5
Smile E1	C-	S U C-	C+	C+	C-
Surprise E2	C+	C+	C-	C+	C+
Disgust E3	C-	C-	S U C+	C+	S U C-
Anger E4	C+	C-	S	S U C-	S
Sadness E5	C-	C+	S	C+	S
Fear E6	C+	S U C+	S U C-	S U C+	SUC+
Neutral E7	S	S	S	S	S

Figure 3.26: Mapping table between characteristic distances and state variables for a given expression.

Due to a lack of data, it has not been possible to derive such heuristic mapping for *Fear*, *Sadness* and *Anger* expressions. Indeed these three expressions are difficult to simulate by

non actor people. For completeness, we included a mapping for these expressions that is derived from MPEG-4 description of the facial features deformations for such expressions.

Figure 3.26 shows that a *Surprise* expression is characterized by the fact that the eyebrows are raised ($V_2 = C^+$), the upper eyelids are opened ($V_1 = C^+$) and the mouth is opened ($V_3 = C^-$ and $V_4 = C^+$). *Smile* expression is characterized by a widely open mouth ($V_3 = V_4 = C^+$) with corners pulled backward to ears ($V_5 = C^-$), slightly closed eyes ($V_1 = C^-$) and slackened eyebrows ($V_2 = S$) or slightly bended eyebrows ($V_2 = C^-$)(noted $V_2 = S \cup C^-$). Note that some symbolic states can take different values for a given expression (for example V_2 for *Smile*). This is justified by the human variability in rendering an emotion in facial expressions.

Figure 3.26 defines the characteristic distances states knowing the expressions. Now we aim at obtaining the expressions knowing the characteristic distances states. For this the mapping of Figure 3.26 can be reformulated as a set of logical rules for each characteristic distance. As an example, Table 3.4 gives the logical rules for D_2: "1" if for the considered facial expression, the corresponding state is reached and "0" otherwise. This table can be interpreted as: if $V_2 = C^-$ then the reached expression corresponds to $E_1 \cup E_3 \cup E_4$ (refinement process).

		E_1	E_2	E_3	E_4	E_5	E_6	E_7
D_2	C^+	0	1	0	0	1	0∪1	0
	S	0∪1	0	0	0	0	0∪1	1
	C^-	0∪1	0	1	1	0	0	0

Table 3.4: Logical rules of symbolic states for characteristic distance D_2 for each expression.

3.7.3 Data fusion process

Human expressions are variable according to the individual. In addition, human is not binary and doubt between several expressions can appear. Moreover, sometimes the emotion is not clearly expressed and then cannot be directly recognized. Finally, based on an automatic segmentation process, errors can appear on the distances. For all these reasons a pure logic system is not sufficient to make a reliable recognition of expressions. These points lead to the choice of a method which is able to model uncertainty and inaccuracy on parameters and on emotions.

We have chosen to use the Transferable Belief Model (TBM) because this approach takes into account the uncertainty of the input information and it allows to explicitly model the doubt between several hypotheses. The TBM have been used in several applications such as image processing, geoscience, medicine, robotic and defense [Val00].

3.7.3.1 Transferable Belief Model

Initially introduced by Dempster [Dem68], it has been revisited by Shafer [Sha76] which showed the interest of this theory as a modeling tools of the uncertainty. Smets ([Sme90], [Sme94]) enriched this theory and defined the Transferable Belief Model (TBM).

The TBM can be seen as a generalization of the theory of probabilities. It considers the frame of discernment $\Omega = \{H_1, ..., H_N\}$ of N exhaustive and exclusive hypotheses characterizing some situations, for example the different expressions of a human face. This means that the solution of the considered problem is unique and that it is obligatorily one of the

hypotheses of Ω. It requires the definition of a Basic Belief Assignment (BBA) that assigns an elementary piece of evidence $m(A)$ to every proposition A of the power set $2^\Omega = \{A/A \subseteq \Omega\} = \{\emptyset, \{H_1\}, \{H_2\}, ..., \{H_N\}, \{H_1, H_2\}, ..., \Omega\}$, that is,

$$\begin{aligned} & m : 2^\Omega \rightarrow [0,1] \\ & A \mapsto m(A), \sum_{A \in 2^\Omega} m(A) = 1 \end{aligned} \tag{3.1}$$

The subset A which contains several hypotheses is called proposition and the subset composed of only one hypothesis H_i is called singleton. By proposition, we mean any disjunction (union) of the hypotheses of the set Ω, including the empty disjunction ($\emptyset$), i.e. none of the hypotheses. For example, the proposition $A = \{H_i, H_j\}$, can be noted $A = H_i \cup H_j$ and means that we consider that either H_i or H_j is true. $m(A)$ corresponds to the piece of evidence of the proposition A. It traduces our belief in the proposition without favoring any of its hypotheses. The propositions whose piece of evidence is not null are called the focal elements. We see that the TBM allows to model doubt in a decision process. Besides, it is well adapted to design a fusion approach where various independent sensors collaborate together to provide more reliable decisions.

In our application the independent sensors correspond to the different characteristic distances which can evolve/move freely one compared to the others due to the motion of the different features (for example the mouth can be opened while the eyes remain stable). The hypotheses H_i correspond to one of the seven facial expressions: Ω_E= {*Smile* (E_1), *Surprise* (E_2), *Disgust* (E_3), *Sadness* (E_4), *Anger* (E_5), *Fear* (E_6), *Neutral* (E_7) }. However, in order to have an intermediate modeling between the numerical values of our characteristic distances D_i and the required expressions, we first define a Basic Belief Assignment related to the characteristic distances states (see section 3.7.3.3 .b). Then, the combination process of the BBAs of all the distance states leads to the definition of the BBAs of the facial expressions (see section 3.7.3.3 .c).

3.7.3.2 Design of Basic Belief Assignment

Using the TBM approach requires the definition of a Basic Belief Assignment to be associated to each independent source of information D_i. It consists in associating a piece of evidence to each proposition A as:

$$\begin{aligned} m_{D_i} : & \\ & A \in 2^\Omega \mapsto m_{D_i}(A) \in [0,1], \\ & \sum_{A \in 2^\Omega} m_{D_i}(A) = 1 \end{aligned} \tag{3.2}$$

where $\Omega = \{C^+, C^-, S\}$ and $2^\Omega = \{\emptyset, \{S\}, \{C^+\}, \{C^-\}, \{S, C^+\}, \{S, C^-\}, \{C^+, C^-\}, \{S, C^+, C^-\}\}$. In our application the frame of discernment 2^Ω is reduced to $\{\{S\}, \{C^+\}, \{C^-\}, \{S, C^+\}, \{S, C^-\}\}$ because the two propositions $\{C^+, C^-\}$ and $\{S, C^+, C^-\}$ are not possible. $\{S, C^+\}$ (resp. $\{S, C^-\}$) corresponds to the doubt states between S and C^+ (resp. S and C^-) and is noted

$S \cup C^+$ (resp. $S \cup C^-$). Similarly and for simplification of notation the propositions $\{S\}$,$\{C^+\}$ and $\{C^-\}$ are noted respectively S,C^+ and C^-.

The piece of evidence $m_{D_i}(V_i)$ associated to each proposition given the characteristic distance D_i is obtained by the function depicted in Figure 3.27. The threshold values (*a, b, c, d, e, f, g, h*) of each model are different and have been derived by statistical analysis on the HCE database for each characteristic distance. The database have been divided into a learning set called HCE_L (13 subjects and 4 expressions, 4680 frames (see section 1.1.4)) and a test set called HCE_T (8 subjects and 4 expressions, 3840 frames). The learning set is then divided into expressive frames noted HCE_{Le} and neutral frames HCE_{Ln}.

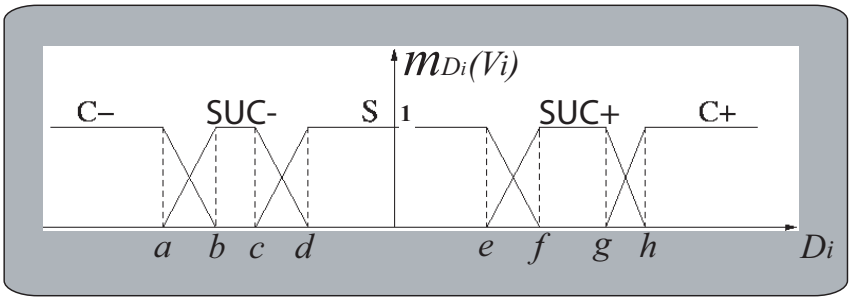

Figure 3.27: Model of basic belief assignment based on characteristic distance D_i for the state variable V_i. For each value of D_i, the sum of the pieces of evidence of the states of D_i is equal to 1.

One model is defined for each characteristic distance independently of the facial expression. The thresholds of each model are defined by statistical analysis. The minimum threshold a is averaged out over the minimum values of the characteristic distances from the HCE_{Le} database (see equation 3.3). Similarly, the maximal threshold h is obtained from the maximum values (see equation 3.4). The middle thresholds d (see equation 3.5) and e (see equation 3.6) are defined respectively as the mean of minimum and maximum, respectively, of the characteristic distances from the HCE_{Ln}.

$$a = mean_{HCE_{Le}}(min(D_i)_{1 \leq i \leq 5}) \tag{3.3}$$
$$h = mean_{HCE_{Le}}(max(D_i)_{1 \leq i \leq 5}) \tag{3.4}$$
$$d = mean_{HCE_{Ln}}(min(D_i)_{1 \leq i \leq 5}) \tag{3.5}$$
$$e = mean_{HCE_{Ln}}(max(D_i)_{1 \leq i \leq 5}) \tag{3.6}$$

The intermediate threshold b (see equation 3.9) is computed as the mean of the characteristic distance values for facial images assigned to the lower state C^- augmented by the median of the maximum values over all the image of the HCE_{Le} database (see equation 3.8). Likewise, the intermediate threshold c (see equation 3.10) is the mean characteristic distance of facial images in state S reduced by the median of the maximum values over all the image

of the HCE_{Le} database (see equation 3.7). The thresholds f and g are obtained similarly.

$$Median_{min} = median_{HCE_{Le}}(min(D_i)_{1\leq i\leq 5}) \tag{3.7}$$
$$Median_{max} = median_{HCE_{Le}}(max(D_i)_{1\leq i\leq 5}) \tag{3.8}$$
$$b = a + Median_{min} \tag{3.9}$$
$$c = h - Median_{max} \tag{3.10}$$

Once these thresholds have been estimated, the Basic Belief Assignment is entirely characterized. The piece of evidence associated to each proposition can be computed and their sum is equal to 1. Figure 3.28 gives examples of the time evolution of pieces of evidence in case of a *Surprise* expression. Let us comment the evolution of the pieces of evidence related to the eyebrows motion characteristic distance D_2 (Figure 3.28.b). At the beginning (from frame 1 to 2), the subject starts with a *Neutral* expression and the piece of evidence $m_{D_2}(S)$ is maximal. Next (starting from frame 3), the subject begins performing a *Surprise* expression, $m_{D_2}(S)$ vanishes while the belief is assigned to $S \cup C^+$ and C^+. Then, the subject eyebrows are fully raised and the piece of evidence $m_{D_2}(C^+)$ is maximum. At the end (starting from frame 16), the subject returns to a *Neutral* expression via the doubt state $S \cup C^+$.

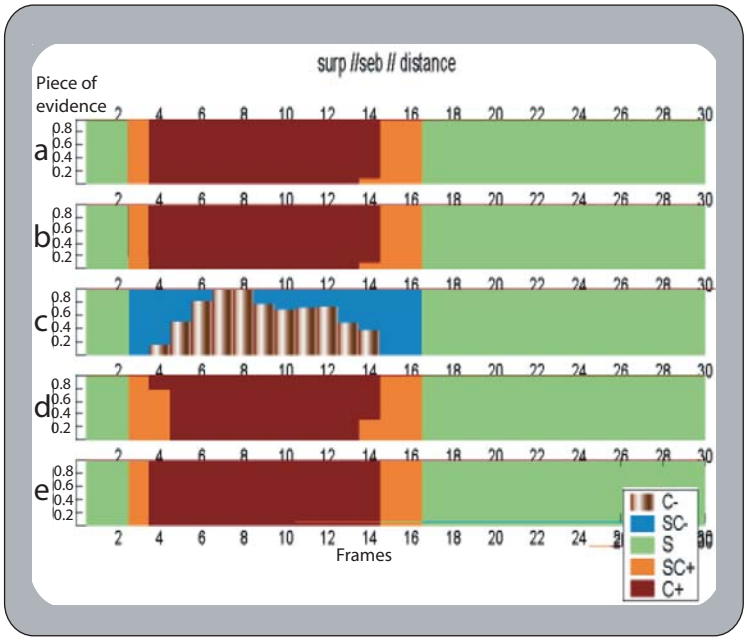

Figure 3.28: Time evolution of pieces of evidence (a) $m_{D_1}(V_1)$, (b) $m_{D_2}(V_2)$, (c) $m_{D_3}(V_3)$, (d) $m_{D_4}(V_4)$ and (e) $m_{D_5}(V_5)$ in case of a *Surprise* expression.

3.7.3.3 Data Fusion and Global Belief Assignment Computation

The salient character of the TBM is the powerful combination operator that allows the integration of information from different sensors. The Basic Belief Assignments described in the previous section can be viewed as independent sensors that score their belief in a proposition given some observations. In order to make the decision about the facial expression, these

sensors are combined to take into account all the available information. To do this, we apply the Dempster combination law (conjunctive combination) [Dem68], [Sme00]. For example we consider two characteristic distances D_i and D_j to which we associate two Basic Belief Assignments m_{D_i} and m_{D_j} defined on the same frame of discernment 2^Ω. Then the joint Basic Belief Assignment $m_{D_{ij}}$ is given using the conjunctive combination (orthogonal sum) as:

$$m_{D_{ij}}(A) = (m_{D_i} \oplus m_{D_j})(A) = \sum_{B \cap C = A} m_{D_i}(B) m_{D_j}(C) \tag{3.11}$$

where A, B and C denote propositions and $B \cap C$ denotes the conjunction (intersection) between the propositions B and C. Obviously, the propositions involved in the joint Basic Belief Assignment are more accurate than the ones in the initial Basic Belief Assignments. Hence, there is less uncertainty in the joint Basic Belief Assignment. The combination is based on the concordance of sensors. It means that the piece of evidence is allocated to the intersections of the propositions which implies that these intersections should not be empty. Then if during one combination, the piece of evidence of the empty set is not null, this implies that there is a conflict between the sensors. It is noted ϕ and its piece of evidence corresponds to:

$$m_{D_{ij}}(\phi) = \sum_{B \cap C = \phi} m_{D_i}(B) m_{D_j}(C) \tag{3.12}$$

In our application, we have a sensor for every characteristic distance with its own frame of discernment. However, we practically need to formulate the joint Basic Belief Assignment in terms of facial expressions. To do so, we use the table of logical rules (see Table 3.4) which allows to associate the piece of evidence of each symbolic state to the corresponding expressions (expressions for which the state is reached) of for all the distances. For example, the Basic Belief Assignment related to D_2 is computed according to its binary table (see Table 3.4) as:

$$m^{\Omega}_{D_2}(C^-) = m^{\Omega_E}_{D_2}(E_1 \cup E_3 \cup E_4) \qquad \text{(see section 3.7.2)}$$

meaning that the piece of evidence (our belief) associated to the state C^- of the characteristic distance D_2 is equal to the piece of evidence (our belief) of the expression E_1 or E_3 or E_4. Similarly, the pieces of evidence of the other expressions can be derived as:

$$m^{\Omega}_{D_2}(S) = m^{\Omega_E}_{D_2}(E_1 \cup E_6 \cup E_7)$$
$$m^{\Omega}_{D_2}(C^+) = m^{\Omega_E}_{D_2}(E_2 \cup E_5 \cup E_6)$$

For the doubt state $S \cup C^+$, S is reached for $E_1 \cup E_6 \cup E_7$ and C^+ is reached for $E_2 \cup E_5 \cup E_6$ hence $S \cup C^+$ is reached for $(E_1 \cup E_6 \cup E_7) \cup (E_2 \cup E_5 \cup E_6) = (E_1 \cup E_2 \cup E_5 \cup E_6 \cup E_7)$.
Then:

$$m^{\Omega}_{D_2}(S \cup C^+) = m^{\Omega_E}_{D_2}(E_1 \cup E_2 \cup E_5 \cup E_6 \cup E_7)$$

Similarly for the doubt state $S \cup C^-$ (see Table 3.4):

$$m^{\Omega}_{D_2}(S \cup C^-) = m^{\Omega_E}_{D_2}(E_1 \cup E_3 \cup E_4 \cup E_6 \cup E_7)$$

The same process is applied for all the characteristic distances using the corresponding logical rules tables. Then for each characteristic distance we obtain the pieces of evidence of the expressions or of doubt between expressions (for example $E_1 \cup E_3 \cup E_4$) which are associated to it.

Once the piece of evidence of the expressions defined for each distance, it is possible to combine them through the use of the orthogonal sum according to all the distances. To be more explicit, consider two Basic Belief Assignments as:

$$\begin{array}{rr} m_{D_i}(E_1) & m_{D_j}(E_1) \\ m_{D_i}(E_2) & m_{D_j}(E_2) \\ m_{D_i}(E_1 \cup E_2) & m_{D_j}(E_2 \cup E_3) \end{array}$$

Their combination leads to the definition of Table 3.5. Each element corresponds to the intersection of expressions associated to D_i and D_j. For example in the second row third column E_2 corresponds to $(E_2) \cap (E_2 \cup E_3)$.

m_{D_i} \ m_{D_j}	E_1	E_2	$E_2 \cup E_3$
E_1	E_1	ϕ	ϕ
E_2	ϕ	E_2	E_2
$E_1 \cup E_2$	E_1	E_2	E_2

Table 3.5: Example of combination of PEs of two distances. ϕ is the empty set.

The piece of evidence of each expression is computed by the orthogonal combination (see Equation 3.11) of results of the two distances:

$$\begin{aligned} m_{D_{ij}}(E_1) &= m_{D_i}(E_1)m_{D_j}(E_1) + m_{D_i}(E_1 \cup E_2)m_{D_j}E_1, \\ m_{D_{ij}}(E_2) &= m_{D_i}(E_2)m_{D_j}(E_2) + m_{D_i}(E_2)m_{D_j}(E_2 \cup E_3) + m_{D_i}(E_1 \cup E_2)m_{D_j}(E_2) \\ &\quad + m_{D_i}(E_1 \cup E_2)m_{D_j}(E_2 \cup E_3), \\ m_{D_{ij}}(\phi) &= m_{D_i}(E_1)m_{D_j}(E_2) + m_{D_i}(E_1)m_{D_j}(E_2 \cup E_3) + m_{D_i}(E_2)m_{D_j}(E_1). \end{aligned}$$

while the initial Basic Belief Assignments are also defined for composite propositions, i.e. in the presence of doubt. Hence, we see that the combination of different sources of information allows reducing or even removing doubt in the decision process.

Note that the empty set can appear and allows handling conflicts between incoherent sensors. The empty set corresponds to situations where the values of characteristic distances

leading to symbolic states configuration do not correspond to any of those defined in Figure 3.26. This has to be related to the fact that Ω_E is not really exhaustive. The underlying facial expression is assigned to *Unknown* expression, noted E_8 in the following.

A decision requires making a choice. However making a choice means taking a risk, except if the result of the combination is perfectly reliable: $m(E_i) = 1$. As it is not always the case, several classical criteria can be used: the plausibility (Pl see equation 3.13 and Figure 3.29 left) which favors the single hypotheses in the case of mixture of expressions, the belief (Bel see equation 3.14 and Figure 3.29 right) which favors the mixture of hypotheses [Sme00] and the pignistic probability (BetP see equation 3.15) which only deals with singleton expressions.

$$\begin{aligned} &Pl : 2^{\Omega} \rightarrow [0,1] \\ &A \rightarrow Pl(A) = \sum_{B \cap A \neq \phi} m(B) \end{aligned} \tag{3.13}$$

Pl(A): plausibility that trueness is in A.

$$\begin{aligned} &Bel : 2^{\Omega} \rightarrow [0,1] \\ &A \rightarrow Bel(A) = \sum_{B \subseteq A, B \neq \phi} m(B) \end{aligned} \tag{3.14}$$

Bel(A): belief that trueness is in A.
A can be either single expressions $A = E_i$ or disjunctions of expressions $A = \cup_i E_i$.

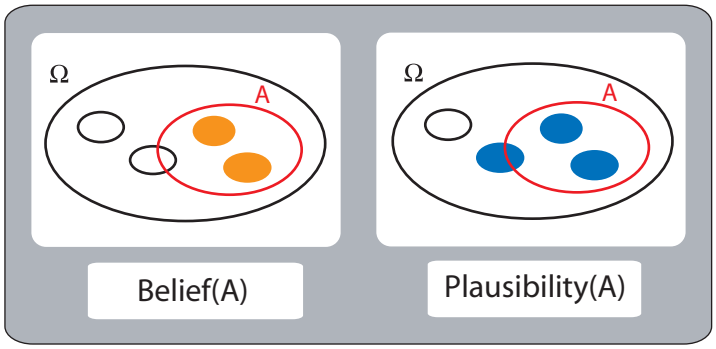

Figure 3.29: Example of belief and plausibility decision process; the sets in color (filled in gray level) are used to compute the piece of evidence of A; on the left the Belief corresponds to the use of sets included in A; on the right the Plausibility corresponds to the use of sets intersected with A.

$$\begin{aligned} &BetP : \Omega \rightarrow [0,1] \\ &B \rightarrow BetP(B) = \sum_{B \in A} \frac{m(A)}{Card(A)} \end{aligned} \tag{3.15}$$

BetP(B): pignistic probability of B.
B is a single expression $B = E_i$.

In our case the basic belief assignment are consonnant then the intersection between the focal elements is not empty. This induces constraints on the choice of the decision criterion. On the one side we want to be able to choose a singleton or a subset of hypotheses. On the other side we want the chosen element to be believable and to be the smallest possible subset.

The classical decision criteria do not allow to reach these objectives. Indeed the pignistic probabilities only allow to choose singleton hypotheses. For the plausibility, in the case of consonnant distributions, all the focal elements have a piece of evidence equal to 1. For the belief, the focal element which has the maximum belief is the one which is the biggest subset.

We choose a compromise taking the proposition maximizing the joint piece of evidence as the decision criterion, that is:

$$A^* = arg(\underset{A \in 2^\Omega}{max}(m_{D_{12345}}(A))) \tag{3.16}$$

For example, if we consider $\Omega = \{E_1, E_2, E_3\}$ and a basic belief assignment:

$$\begin{aligned} m(E_1) &= 0.1 \\ m(E_1 \cup E_2) &= 0.2 \\ m(E_1 \cup E_3) &= 0.5 \\ m(E_1 \cup E_2 \cup E_3) &= 0.2 \end{aligned}$$

We choose $E_1 \cup E_3$ which is a compromise between $E_1 \cup E_2 \cup E_3$ the most believable subset and E_1 the most precise choice but not very believable.

In future works it will be necessary to study more deeply the definition of decision criterion compatible with the consonnant nature of this type of distribution.

3.7.4 Post processing

The results given by the fusion process are relatively satisfactory. They will be commented in section 3.8. Meanwhile, some pairs of expressions are difficult to discriminate, for instance *Disgust-Smile* or *Fear-Surprise*.

As explained in section 2, the main idea of our approach is that the use of the five characteristic distances is sufficient for facial expressions recognition. Thus our modeling process based on the TBM only integrates the characteristic distances states. The analysis (see section 3.8) shows that they are necessary to dissociate between the studied facial expressions. However it appears that they are not sufficient to dissociate between some confused expressions. In the case of *Smile* and *Disgust*, we use two other parameters (the presence of nasal root wrinkles and mouth shape ratio) as additional information in the case of this confusion. Indeed these two features are not necessary to characterize all the studied expressions but they can be useful to dissociate specifically these two ones. In the fusion process they only have to be added in a post-processing step when this confusion appears.

Figure 3.30 shows examples of nasal roots and mouth shapes in case of *Disgust* or *Smile*. Wrinkles appear in the nasal root in the case of *Disgust* (see Figure 3.30.a) contrary to the case of *Smile* where they are absent (see Figure 3.30.b). Moreover the mouth shape in the case of *Disgust* (see Figure 3.30.c) is different from its shape in the case of *Smile* (see Figure 3.30.d).

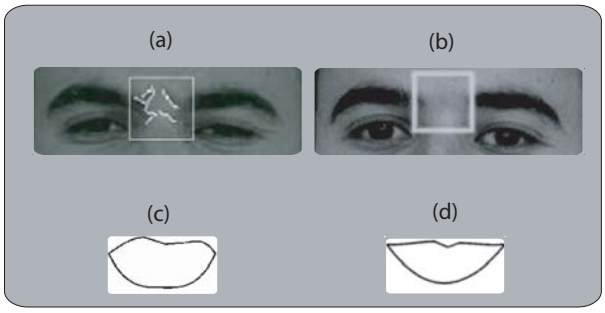

Figure 3.30: Nasal root (a) with wrinkles in a *Disgust* facial image and (b) without wrinkles in a *Smile* facial image. Mouth shape in case of (c) *Disgust* and (d) *Smile*.

Nasal root wrinkles are detected in facial images by using a Canny edge detector [Can86]. The presence or absence of wrinkles is actually decided by comparing the number of edge points in the nasal root in the current expressive image with the number of edge points in the nasal root of a *Neutral* facial image. If there are about twice more edge points in the current image than in the reference image, wrinkles are supposed to be present. The Canny edge threshold is set by expertise. We take a high threshold to minimize the risk of errors. Then the system keeps the doubt instead of taking the risk of making a wrong decision.

For the mouth shape ratio, D_3/D_4 is considered. It can be larger or smaller than its corresponding value in the *Neutral* expression.

At this stage, the rules are strictly logical and are described in the following algorithm:

1: **if** nasal root wrinkles are present **then**
2: *Disgust* is chosen
3: **else if** ratio between mouth width and mouth height is higher than its value in the Neutral state **then**
4: *Smile* is chosen
5: **else**
6: *Disgust* is chosen
7: **end if**

3.8 Results

3.8.1 Implementation

One characteristic of the belief theory is the risk of the combinatory explosion. This problem is bypassed with a programming technique that consists in coding facial expressions and their combination in using binary words. The intersection is then realized by logical AND [Yag94]. Table 3.6 gives an illustration of this implementation for 3 expressions.

The code of intersection for two propositions corresponds to a logical AND between the two following logical codes:

$$(E_1 \cup E_2) \cap (E_1 \cup E_3) = 011 \cap 101 = 001 = E_1 \tag{3.17}$$

E_3	E_2	E_1	E
0	0	1	E_1
0	1	0	E_2
1	0	0	E_3
0	1	1	$E_1 \cup E_2$
1	0	1	$E_1 \cup E_3$
1	1	1	$E_1 \cup E_2 \cup E_3$

Table 3.6: Example of binary code for 3 expressions.

3.8.2 Expertise and test database

The expertise is based on the HCE database. The database is divided into a learning set HCE_L (13 subjects and 4 expressions) and a test set HCE_T (8 subjects and 4 expressions). Three databases are used for the test step: the HCE_T database; the CKE database (30 subjects have been chosen for *Smile*, 25 for *Surprise*, 17 for *Disgust*, _ for *Sadness*, _ for *Fear*, _ for *Anger* and 72 for *Neutral*) and the DCE database (7 expressions with 8 females and 8 males) (see section 1.1).

For the CKE and DCE databases only manual data can be used. Indeed as it is based on the chrominance information, automatic segmentation of the mouth is not possible on these databases (gray level frames) and facial expressions classification can only be evaluated based on data obtained from manual segmentation. In addition from each one of the selected sequences of the CKE database, only two images have been used: the *neutral* state and the *apex* of the expression (it corresponds to the maximum of intensity reached by the facial expression associated with a particular distances states configuration).

3.8.3 Classification rates analysis

We propose an analysis of the classification rates based on two types of input data for the TBM:

- a classification based on distances obtained from a manual segmentation of eyes, eyebrows and mouth. This is used to validate the data fusion process alone.
- a classification based on distances obtained from our automatic segmentation algorithm (see chapter 2).

The aim is to identify the classification errors due to the automatic segmentation and to dissociate them from errors due to the facial expressions classification modeling itself.

The classification rates are obtained on two sets of expressions: a classification on 4 expressions (*Smile*, *Surprise*, *Disgust*, *Neutral*) and a classification on 7 expressions (*Smile*, *Surprise*, *Disgust*, *Anger*, *Sadness*, *Fear*, *Neutral*). Indeed our analysis of expressions and the definition of the rules table (Figure 3.26) has been carried out on the HCE_L database. As explained before it has only been possible to derive rules for the 3 first expressions (*Smile*, *Surprise*, *Disgust*), the rules for the 3 following ones (*Anger*, *Sadness*, *Fear*) have been taken from MPEG4 description. The classification on the 4 expressions allows to evaluate the

behavior of our classification system based on our defined rules. The classification on the 7 expressions allows to evaluate the correctness of the 3 added MPEG4 rules and their influence on the behavior of our classification system.

3.8.3.1 Results on data obtained from manual segmentation

Results on the HCE_T database

The performances of the resulting classification system are evaluated on the HCE_T database and results on 4 expressions are given in Table 3.7. Columns correspond to the facial expressions labeled by an expert and rows correspond to classification rates obtained by the system. The first rows correspond to the single propositions (singletons), including the *Unknown* expression E_8, the next two rows correspond to pair propositions, i.e presence of doubt, and the last row includes all the other possible propositions (doubt between more than 2 expressions).

We observe that good classification rates are obtained on the expressions E_1 (*Smile*), E_2 (*Surprise*) and E_7 (*Neutral*). On the contrary, the classification rate for expression E_3 (*Disgust*) is lower. This can be explained by the high variability of this expression between subjects (see Figure 3.31 bottom) and the difficulty for a non-actor person to simulate this expression (see Figure 3.31 top).

Expert / System	E_1	E_2	E_3	E_7
E_1 *Smile*	<u>76.36</u>	0	9.48	0
E_2 *Surprise*	0	<u>85.04</u>	0	0
E_3 *Disgust*	0	0	<u>43.10</u>	0
E_7 *Neutral*	6.66	0.80	15.51	<u>88</u>
E_8 *Unknown*	6.06	11.80	12.06	0
$E_1 \cup E_3$	<u>10.90</u>	0	<u>8.62</u>	0
others	0.02	2.36	11.23	12
Total	87.26	85.04	51.72	88

Table 3.7: Classification rates in percent on 4 expressions with data obtained from manual segmentation on the HCE_T database. The Total row corresponds to the classification rates obtained by summing the underlined results of each corresponding column.

We also observe that the post-processing step does not yield to the total cancellation of doubt state between *Smile* and *Disgust* (see section 3.7.4). The system has the highest belief in the disjunction of both propositions but it cannot discriminate between them. This has to be related to Figure 3.26 where states variables for *Smile* and *Disgust* can take the same values depending on the values of the characteristic distances. Nevertheless the post-processing step allows a significant increase of the classification results. Table 3.8 gives the comparison results with and without the post-processing step. The recognition rate for E_1 (*Smile*) increases by 15% and $E_1 \cup E_3$ (*Smile-Disgust*) decreases by 17% (2% of false detection of *Disgust*). E_3 (*Disgust*) increases by 11% and $E_1 \cup E_3$ (*Smile-Disgust*) decreases by 19% (8% of false detection of *Smile*).

Figure 3.31: Examples of facial images in case of *Disgust* expression: first row, poor simulation by non-actor subjects and second row high variability between subjects.

System \ Expert		E_1	E_3
without	E_1 *Smile*	<u>66.4</u>	8.77
	E_3 *Disgust*	0	<u>38.82</u>
	$E_1 \cup E_3$	13.13	10.64
with	E_1 *Smile*	<u>76.36</u>	9.48
	E_3 *Disgust*	2	<u>43.10</u>
	$E_1 \cup E_3$	10.90	8.62

Table 3.8: Classification on the HCE_T database with and without the use of the post processing step.

Given the fact that the doubt state *Smile-Disgust* is related to the rules defined in Figure 3.26, it is not due to classification errors of the proposed system. Moreover in this case, the system is sure that the current expression is one of these two ones and that it is not one of the others. It is thus possible to consider it as a good classification and to associate it to the corresponding expression. This allows us to add their respecting rates leading to the results of the last row of Table 3.7 called Total. In Table 3.7 and in all the following rates tables, row Total is obtained summing the underlined rates corresponding to the expression and the doubt state where this expression appears.

The performances of the resulting classification system are also evaluated on 7 expressions (see Table 3.9). We can observe that the classification rates of *Smile* and *Disgust* have not changed. This means that their classification rules specify well these two expressions relatively to the others. The classification based on the 7 expressions introduces another source of confusion between E_2 (*Surprise*) and E_6 (*Fear*). This is due to the fact that these two expressions are hard to distinguish using only characteristic distances, even for human experts (see Figure 3.32). Then this doubt state is not considered as a failure of the

classification system, on the contrary it could be preferable to keep the doubt between these two expressions instead of taking the risk of choosing the wrong expression. The TBM are actually well adapted for such a scenario.

System \ Expert	E_1	E_2	E_3	E_7
E_1 *Smile*	<u>76.36</u>	0	9.48	0
E_2 *Surprise*	0	<u>12.6</u>	0	0
E_3 *Disgust*	0	0	<u>43.10</u>	0
E_4 *Anger*	0	0	0	0
E_5 *Sadness*	0	0	0	0
E_6 *Fear*	0	0	0	0
E_7 *Neutral*	6.66	0.8	15.51	<u>88</u>
E_8 *Unknown*	6.06	11.80	12.06	0
$E_1 \cup E_3$	<u>10.90</u>	0	<u>8.62</u>	0
$E_2 \cup E_6$	0	<u>72.44</u>	0	0
others	0.02	2.36	11.23	12
Total	**87.26**	**85.04**	**51.72**	**88**

Table 3.9: Classification rates in percent on 7 expressions with data obtained from manual segmentation on the HCE_T database. The Total row corresponds to the classification rates obtained by summing the underlined results of each corresponding column.

Similarly to the doubt state *Smile-Disgust* in Table 3.7, given the fact that the doubt state *Surprise-Fear* is related to the rules defined in Figure 3.26, it is possible to consider it as a good classification and to associate it to the corresponding expression allowing us to add their respecting rates leading to the results of the last row of Table 3.9 (row Total).

Figure 3.32: Examples of confusing images: left, *Surprise* expression and right *Fear* expression.

Finally, some images are recognized as *Unknown*. They typically correspond to intermediate images where the subject is neither in *Neutral* state nor in a particular expression. Figure 3.33 shows three images as they appear in a recording of *Smile* expression.

Figures 3.34, 3.35 and 3.36 present visual examples of classification on the 7 expressions for various subjects showing the 4 expressions (*Disgust*, *Smile*, *Surprise* and *Neutral*). The examples are presented in three columns per line which correspond respectively to the initial *Neutral* state, the beginning of the expression and the apex of the expression. These examples

Figure 3.33: Example of a *Neutral* image (left) followed by an *Unknown* image (middle) and a *Smile* image (right).

confirm that classification based on the highest piece of evidence is very often correct when considering the *Neutral* state and the apex of the expression.

In Figure 3.34, the intermediate frames correspond to the intermediate states between *Neutral* and *Disgust* apex. In Figures 3.34.b.3 and 3.34.c.3, we notice the sensitivity of the system to recognize *Disgust* at different intensities. Figures 3.34.b.2 and 3.34.b.4 show two states of doubt between *Smile* and *Disgust*. These examples are hardly distinguished even by a human expert. Figure 3.34.b.1 shows the *Unknown* state which corresponds to intermediate state between *Neutral* and *Disgust* expression.

Figure 3.35 shows the result of *Smile* classification. Figures 3.35.2, 3.35.3, 3.35.5 and 3.35.6 show the sensitivity of the system to different intensities of *Smile* expression. In Figure 3.35.b.1, the system classifies the intermediate state as *Unknown*.

In Figure 3.36, we see the difficulty to separate *Surprise* and *Fear*. However the system is completely sure that it is one of the two expressions and not any other. This incapacity to distinguish between these two expressions is confirmed by human expertise. The only way to separate them is to add information about the context or information from another modality such as speech signal for example.

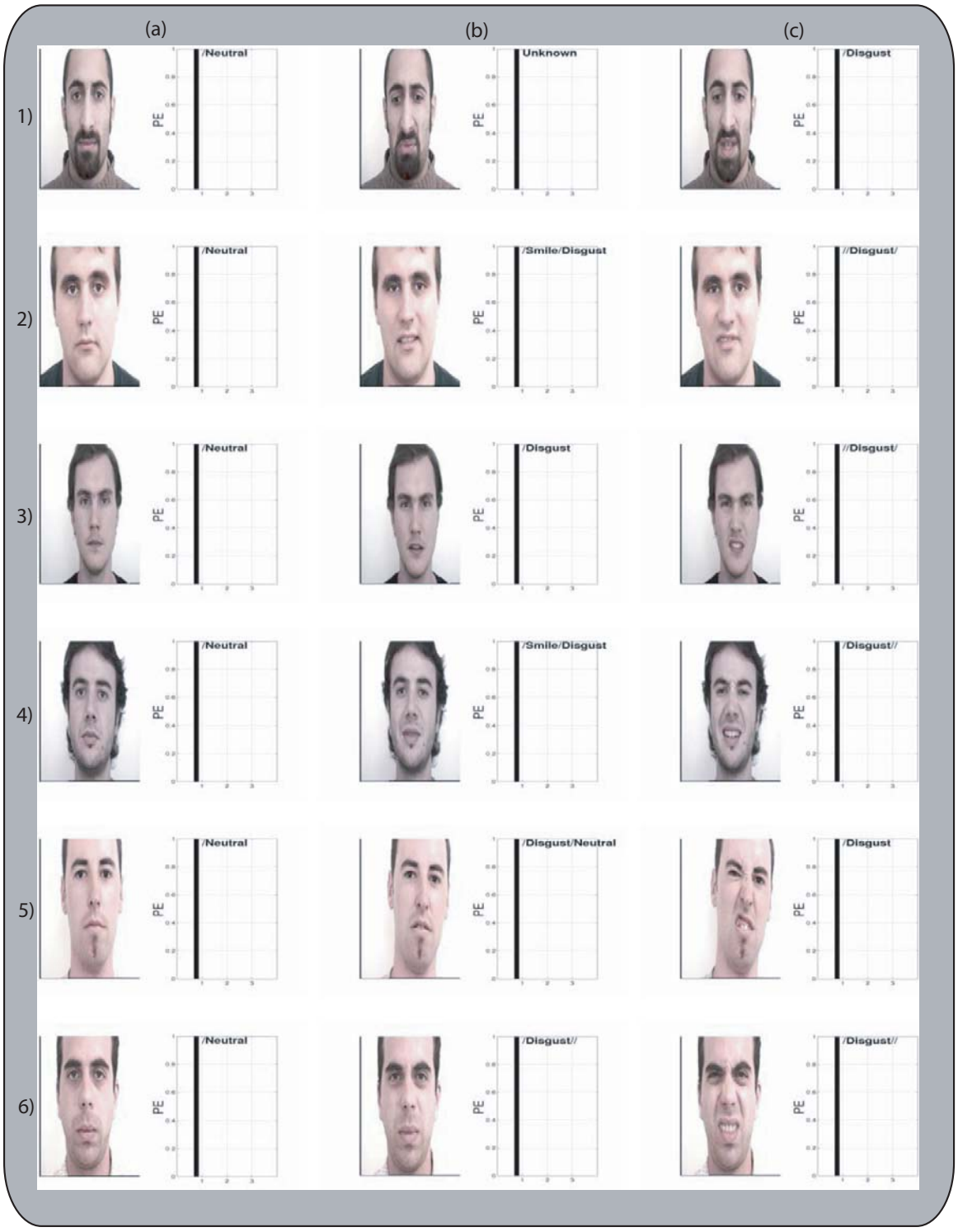

Figure 3.34: Examples of *Disgust* facial expressions: (a) initial *Disgust* state, (b) transition to *Disgust* and (c) apex of *Disgust*. Bar graphs show the piece of evidence for the recognized expression.

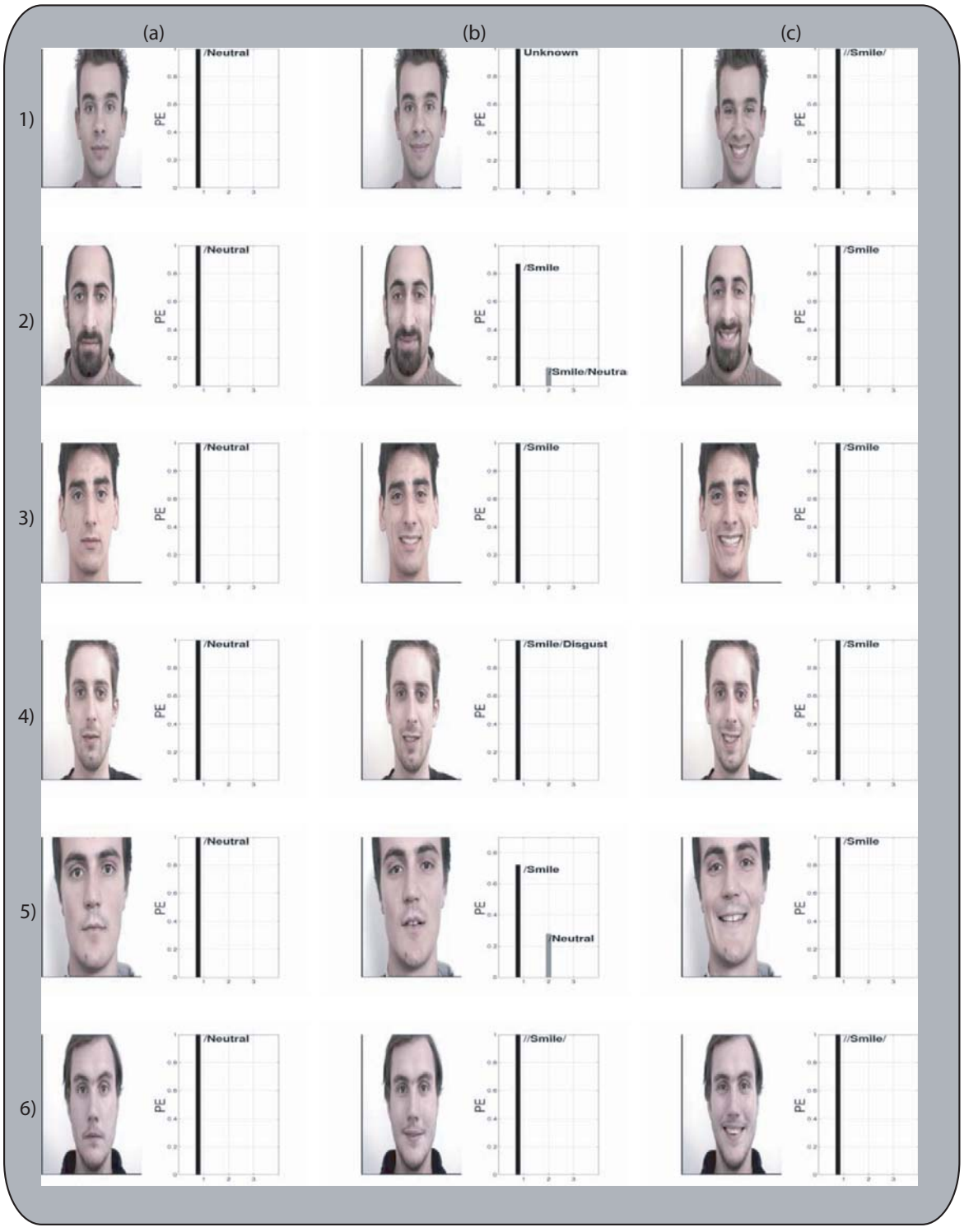

Figure 3.35: Examples of *Smile* facial expressions: (a) initial *Neutral* state, (b) transition to *Smile* and (c) apex of *Smile*. Bar graphs show the piece of evidence for the recognized expression.

Figure 3.36: Examples of *Smile* facial expressions: (a) initial *Neutral* state, (b) transition to *Surprise* and (c) apex of *Surprise*. Bar graphs show the piece of evidence for the recognized expression .

In order to evaluate the robustness of the proposed recognition system to different variations (gender, ethnicity, difference of expressions, etc), the system is also tested on the CKE database and on the DCE database.

Results on the CKE database

If we consider the singleton results, the classification rates of Table 3.10 for this database are comparable with those of Table 3.9. However, if we take into account the doubt between expressions (the last row of Table 3.10) we can observe that the classification rates are better than the results obtained on the HCE database. This is due to the fact that the CKE database is composed of very acted and similar expressions.

System \ Expert	E_1	E_2	E_3	E_7
E_1 *Smile*	64.51	0	0	0
E_2 *Surprise*	0	16	0	0
E_3 *Disgust*	0	0	52.94	0
E_7 *Neutral*	0	0	0	100
E_8 *Unknown*	3.22	0	0	0
$E_1 \cup E_3$	32.27	0	47.05	0
$E_2 \cup E_6$	0	84	0	0
others			0.01	
Total	**96.76**	**100**	**99.99**	**100**

Table 3.10: Classification rates in percent of the system on data obtained from manual segmentation on the CKE database.

Results on the DCE database

Similarly to the CKE database, the system gives good classification rates on the DCE database and the conclusions are the same as those made on the CKE database.

System \ Expert	E_1	E_2	E_3	E_7
E_1 *Smile*	62.50	0	0	0
E_2 *Surprise*	0	25	0	0
E_3 *Disgust*	0	0	75	0
E_7 *Neutral*	0	0	0	100
E_8 *Unknown*	0	0	25	0
$E_1 \cup E_3$	37.50	0	0	0
$E_2 \cup E_6$	0	75	0	0
Total	**100**	**100**	**75**	**100**

Table 3.11: Classification rates in percent on data obtained from manual segmentation on the DCE database.

Figure 3.37 shows examples of classification results on images of the CKE and the DCE databases.

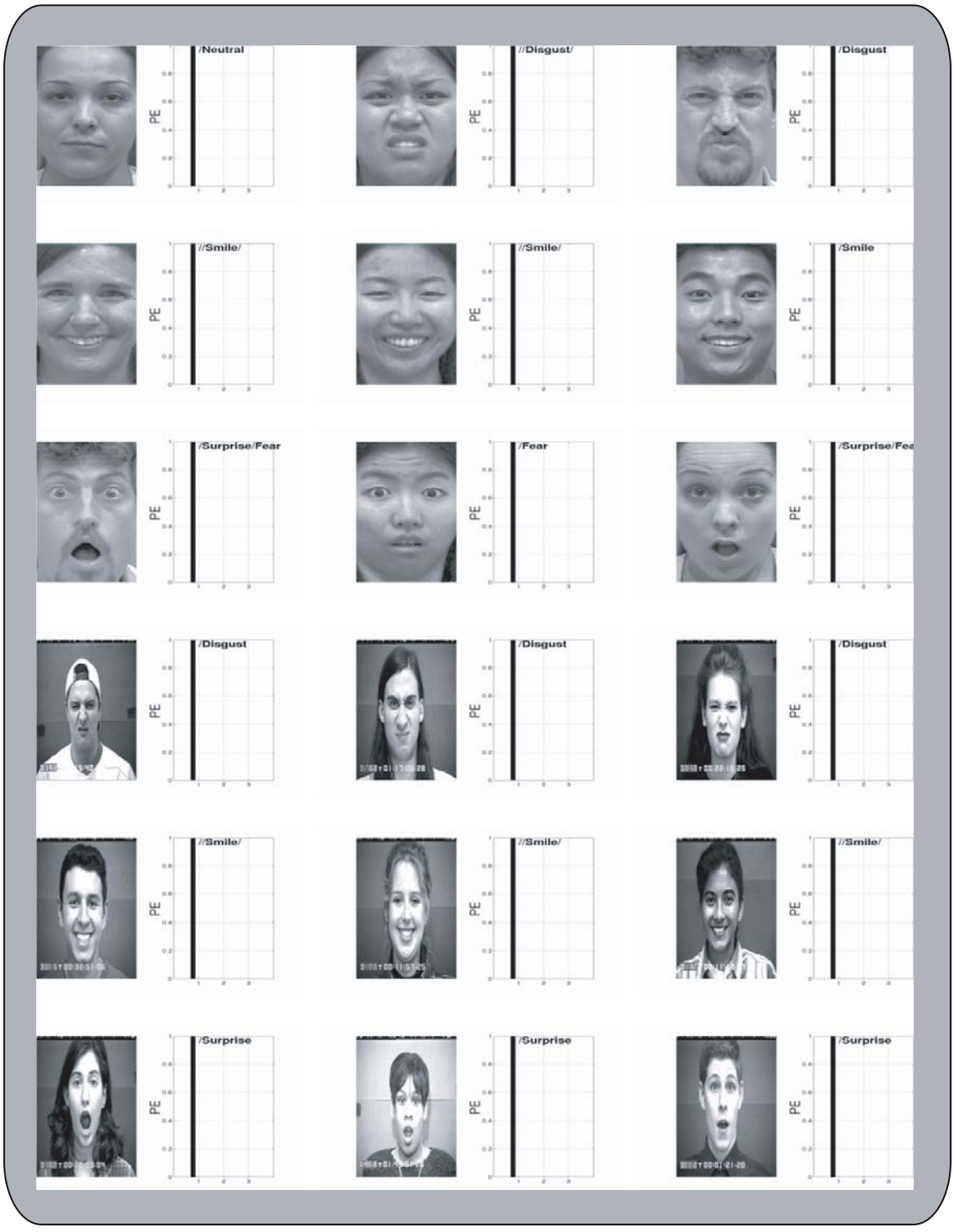

Figure 3.37: Examples of classification of facial expressions: row 1,2,3 shows images from the DCE database and row 4,5,6 shows images from the CKE database.

Contrary to the HCE database, these two databases present also examples of the three remaining universal emotional expressions, namely *Fear*, *Anger* and *Sadness*. Table 3.12 gives the classification results on these three facial expressions obtained on the two databases so as to validate the rules presented in Figure 3.26 and defined using MPEG-4 rules (see Section 3.7.2). Good rates are obtained for *Fear* and *Anger* expression. We also observe the same confusion for *Fear* and *Surprise*. Bad rates are obtained for *Sadness*. The main reason is that, similarly to *Disgust*, this expression is difficult to simulate by non-actor people (this observation is also reported in [Yac96], [Ros96] and [Bla97]). Secondly the MPEG-4 rule for *Sadness* recognition may not be well-defined for the used databases or requires other information to be well classified.

System \ Expert	E_4	E_5	E_6
E_4 *Anger*	79	0	0
E_5 *Sadness*	0	49	0
E_6 *Fear*	0	0	59
E_8 *Unknown*	21	51	29
$E_2 \cup E_6$	0	0	14
others	0	0	0
Total	**79**	**49**	**71**

Table 3.12: Mean classification rates in percent of the CKE and the DCE databases on *Anger*, *Sadness* and *Fear* expressions.

3.8.3.2 Results with data obtained from the automatic segmentation

In this section our aim is to evaluate the robustness of the classification system to occurring segmentation errors due to our facial features segmentation system (see section2).

For comparison purpose, the classification results in the same conditions are also given considering 4 and 7 expressions.

Table 3.13 gives the classification rates on 4 expressions based on the data obtained from our segmentation algorithm. Only results obtained on the HCE_T database are reported because mouth segmentation is based on chrominance information then it has not been possible to produce automatic segmentation on the CKE and DCE databases which are in gray level.

The classification rates based on our segmentation appear to be lower than those based on the manual segmentation (see Table 3.9). This result was expected while comparing the precision between the automatic and the manual segmentation of eyes, eyebrows and mouth characteristic points (see section 2.7.6). However, the most interesting observation is the distribution of the new classification rates.

For *Smile* expression the classification rates decrease mainly in favor of the doubt between *Smile* and *Disgust*. This is due to the imprecision of the automatic segmentation compared to the manual segmentation. The interesting observation is that the system does not traduce this imprecision into classification errors but into doubt between the two expressions.

The classification rates of *Disgust* are very comparable to those obtained with the manual segmentation but their distribution is different. The segmentation imprecision is traduced by the increase of doubt state *Smile-Disgust*. In addition this expression is often classified as

System \ Expert	E_1	E_2	E_3	E_7
E_1 *Smile*	<u>26</u>	0	2.80	1.32
E_2 *Surprise*	0	<u>68</u>	0	0
E_3 *Disgust*	2	0	<u>33</u>	0
E_7 *Neutral*	3.60	2.50	2.20	73
E_8 *Unknown*	10	27	45	7.24
$E_1 \cup E_3$	<u>54</u>	0	<u>16</u>	5.26
others	4.4	2.50	1	13.18
Total	**80**	**68**	**49**	**73**

Table 3.13: Classification rates in percent on 4 expressions on data obtained from our automatic segmentation on the HCE_T database.

Unknown. This is mainly due to the fact that this expression is difficult to simulate by non actor people. Consequently the simulated facial features deformations do not always correspond to the *Disgust* expression.

For the *Neutral* expression, the analysis of the whole set of sequences shows that these results are due to the dispersion of the eyes, eyebrows and mouth key points segmentation in the case of intermediate states between *Neutral* and the considered expression.

In order to better evaluate the suitability of the TBM on data obtained from automatic segmentation, we also test the robustness of our classification in the case of 7 expressions (see Table 3.14).

System \ Expert	E_1	E_2	E_3	E_7
E_1 *Smile*	<u>16</u>	0	0	0.7
E_2 *Surprise*	0	<u>0</u>	0	0
E_3 *Disgust*	6	2	<u>43</u>	0
E_4 *Anger*	0	0	0	0
E_5 *Sadness*	1.20	0	1.10	1.9
E_6 *Fear*	0	24	0	0
E_7 *Neutral*	1.80	1.30	5	68
E_8 *Unknown*	9	9.40	44	4.80
$E_1 \cup E_3$	<u>56</u>	0	<u>4</u>	2.6
$E_2 \cup E_6$	0	<u>61</u>	0	0
others	10	2.30	2.90	22
Total	**72**	**61**	**47**	**68**

Table 3.14: Classification rates in percent on 7 expressions on data obtained from our automatic segmentation on the HCE_T database.

The classification rates based on our automatic segmentation appear to be lower than those based on the manual segmentation (see Table 3.9).

Similarly to the observations done on Table 3.13, the classification rates of *Smile* expression decreases mainly in favor of the doubt between *Smile* and *Disgust*.

For the *Disgust* expression, the classification rates do not change while the *Unknown* ones increase for the same reason as explained in the comments of Table 3.13.

The same observations can be done for *Smile* and *Disgust* as the ones done on Table 3.13.

The classification rates of *Surprise* decrease in favor of *Fear* expression. However it has to be stressed that these two expressions are difficult to distinguish even by human observer (see figure 3.32).

Finally we can notice that the classification rates of the false detections corresponding to doubt between more than three expressions (noted *others*) for all the considered expressions are quite low ($\leq 9\%$). Then the imprecision of the segmentation results are not traduced as classification errors but mainly as doubt between confused expressions.

These results show that the TBM are very suitable for our classification problem and allow to deal with imprecise data and to be robust to the dispersion of the automatic segmentation results.

However compared with results obtained on manual segmentation, these results also show that some classification errors are only due to segmentation errors. For example, we can observe the appearance of *Sadness* expression in the classification of *Disgust* expression which is caused by punctual false detections of the interior eyebrows corners leading to errors on the estimation of the characteristic distance D2. In the case of *Surprise*, punctual false detections of the mouth corners errors on the estimation of the characteristic distance D3 lead to the choice of *Fear* expression instead of *Surprise* expression. Indeed the TBM allows to deal with imprecise data but can not correct the false ones. Then additional information needs to be added to make the classification system able to handle with these segmentation errors. To do this the next step of our modeling is the introduction of the temporal information in the classification process.

3.9 Conclusion

Based on manual data, this rule-based method proves to be well adapted to the problem of facial expressions classification. Indeed it allows to handle with doubt between expressions (for example *Smile* or *Disgust*, *Surprise* or *Fear*) instead of forcing the recognition of a wrong expression. Besides, in the presence of doubt between two expressions, it is sometimes preferable to consider that both of them are possible rather than taking the risk of choosing only one.

In addition, one of the interesting characteristic of the use of the TBM is its ability to model unknown expressions corresponding to all the configurations of distances states unknown to the system. In other classifiers such Bayesian classifier or HMM, it corresponds to a new expression that belongs to a finite set of expressions added to the already defined ones (see Appendix 7.1). Obviously, this new expression does not contain all the possible facial configurations and some *unknown* ones can then be misclassified. This is not the case with the use of the TBM which directly affects new configurations to the *Unknown* expression.

The main goal of our work was the validation of our hypothesis that the permanent facial features contours are sufficient for the recognition of facial expressions. Results obtained on

manual segmentation have shown that the use of characteristic distances computed on facial skeletons is necessary to dissociate between facial expressions. However some of them remain difficult to distinguish (*Smile* and *Disgust*, *Surprise* and *Fear*). Then the use of characteristic distances is not sufficient to dissociate between them and additional information is needed to improve the classification performances (for example the shape of the feature contours or more global information coming from a statistical analysis of the whole face [Buc06]).

Based on automatic segmentation data, classification results are lower than those obtained on manual data. However the TBM proved to be robust to these imprecise data, increasing doubt states between confused expression rather than choosing the wrong one.

To go further it is interesting to observe that in daily life a facial expression is not static but occurs inside a temporal sequence. Indeed, a facial expression is the result of dynamic and progressive combinations of facial features deformations which are not always synchronous (see Figure 3.38). In this case the expression can only be recognized at its apex (third frame in Figure 3.38). However there is no mean to identify the apex of the expression in a sequence. Then the only way to recognize the considered expression is to take into account all the facial features deformations which requires to take into account temporal evolution, beyond static classification.

Figure 3.38: Example of surprise expression. From left to right: *Neutral* state; opening of the eyes; opening of the eyes and the mouth (apex of the expression); slackened eyes and open mouth; *Neutral* state

Then a first improvement of our classification system consists in introducing a temporal information toward a dynamic classification system of facial expression sequences. Next chapter presents our preliminary works on this matter.

Chapter 4

Classification based on dynamic data

From a psychological point of view, it has been shown by Bassili [Bas78], that facial expressions can be more accurately recognized from image sequences than from single images. His experiments used point-light conditions, i.e. subjects viewed image sequences in which only white dots on a darkened surface of the face were visible. Recognition rates on image sequences gave best results than those performed on static images.

Based on these observations some authors have recently worked to model the dynamic information of the facial features deformations on the facial expressions classification. Cohen *et al* [Coh03b] proposed a multi-level HMMs framework (see Figure 4.1) to perform an automatic segmentation and recognition of facial expressions sequences. The first level of the architecture is composed of independent HMMs related to the six universal expressions. At each frame, the motion feature [Coh03a] according to the *Neutral* state is used as the input of the six expression-specific HMMs. Their outputs are used as an observation vector to the high-level HMMs. This latter consists in seven states, one for each of the six expressions and one for *Neutral* (see Figure 4.1). The transitions between expressions are imposed to pass by the *Neutral* state. This modeling allows to obtain the probability of the sequence in displaying one expression and in not displaying any of the other ones. The recognition is done by decoding at each time the state of the high-level HMMs since the state corresponds to the displayed expression in the video sequence at this time. Then each expression sequence contains several instances of each expression with *Neutral* instance separating between them.

Busso [Bus04] proposes a system based on three cameras (see Figure 4.2 left) and 102 markers (see Figure 4.2 right). Dynamic classification of expression sequences is based on the combination of the static classification results obtained on the whole set of frames of the studied sequence. The studied expressions are *Sadness*, *Anger*, *Happiness* and *Neutral*. Each frame of the sequence is divided into five blocks: forehead, eyebrows, low eye, right cheek and left cheek area (see Figure 4.2 right). For each block, the 3D coordinates of markers in the block are concatenated together to form a data vector. Then Principal Component Analysis (PCA) method is applied to each vector to obtain for each one a 10-dimensional vector. For each frame and for each one of the 5 blocks, the 10-dimensional features were classified using a K-nearest neighbor (KNN) classifier. Then for each sequence and for each block the number of apparition of each expression is counted, obtaining a 4-dimensional vector for each block. These latter are added to form a single vector and a SVM classifier was implemented to classify it into one of the 4 studied expressions.

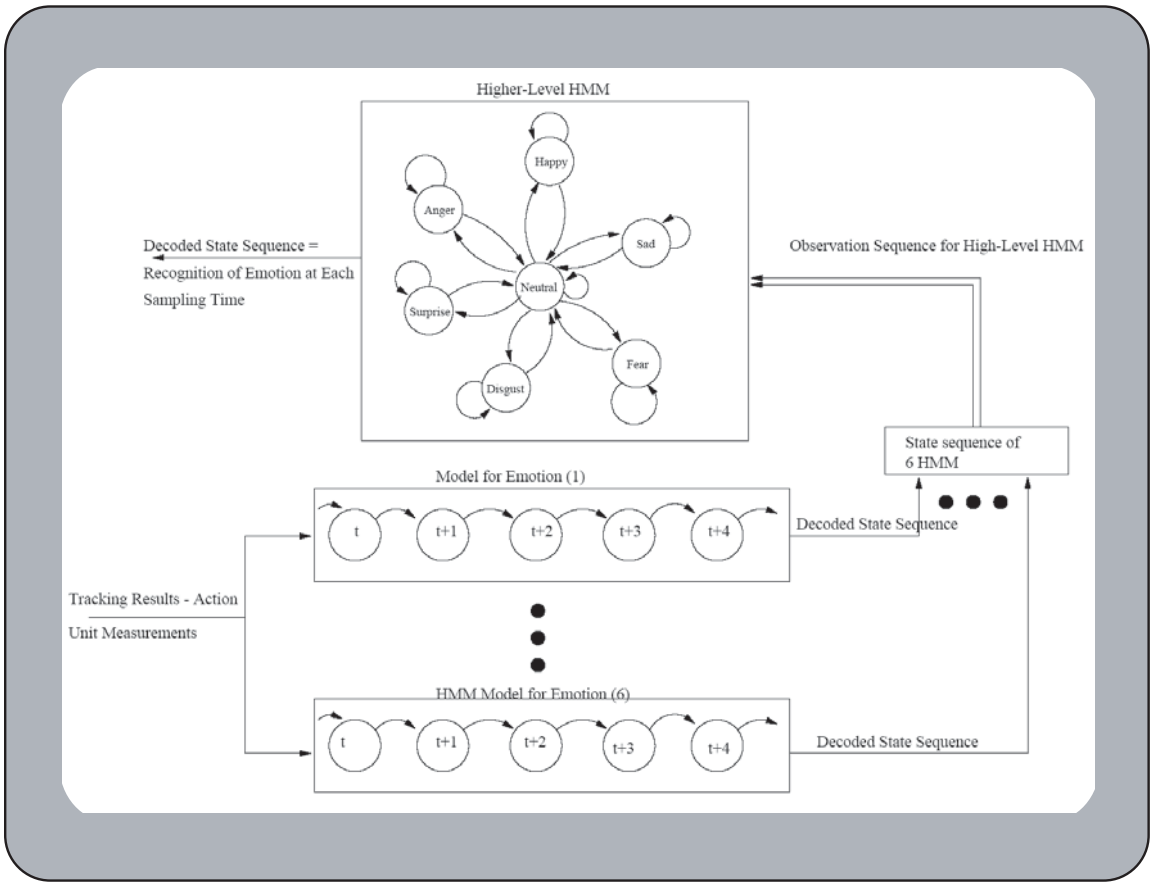

Figure 4.1: Multilevel HMMs architecture for the dynamic recognition of emotion [Coh03b].

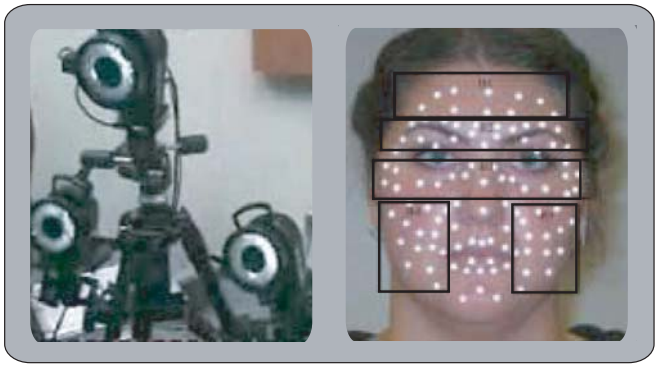

Figure 4.2: Left: three used cameras; right: marked points and facial areas [Bus04].

Zhang Y. and Qiang J. [Zha05] propose a multi-sensory information fusion technique based on a dynamic Bayesian network (DBN). The modeling of the temporal information consists in adding at each time t the classification result obtained at time $t-1$ to the characteristic features vector. The used features correspond to geometrical relationships from the permanent facial features (eyes, eyebrows, nose and mouth) and transient features (see Figure 4.3) which are used for Action Units detection.

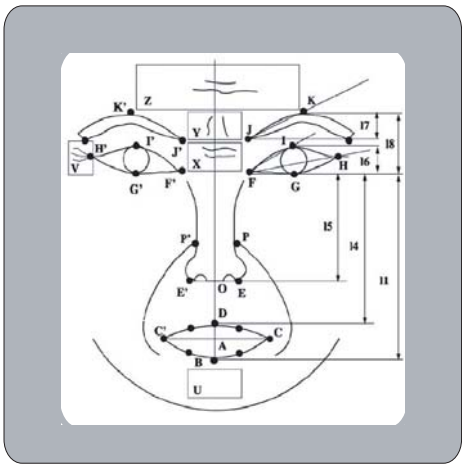

Figure 4.3: Geometrical relationships of facial feature points where the rectangles represent the regions of furrows and wrinkles [Zha05].

The permanent facial features are automatically detected in the first frame and tracked in the remaining frames of the sequence. At each frame of the sequence they combine the classification results of the AUs obtained by the DBN using Ekman FACS [Ekm78] in order to model the dynamic deformations of facial expressions. The fusion of the probability results of all AUs allows to obtain the probability result of the corresponding facial expressions. Some visual classification results are presented on 4 sequences involving multiple expressions. However, they do not give any explicit classification rates.

Pantic and Patras ([Pan05a], [Pan06]) propose a method for recognizing temporal segments (beginning, apex, ending) of facial action units (AUs) during a facial expression sequence. 20 facial frontal feature points (see Figure 4.4 right) and 15 facial profile points (see Figure 4.4 left) initialized in the first frame are tracked in the remaining frames of the sequence. Facial deformations are coded in AUs and are divided into three segments: the onset (beginning), the apex (peak), and the offset (ending). According to the profile view based method or the frontal view based method, a rule based method is defined to uniquely encode the temporal segments of 27 AUs or combination of AUs over each five consecutive frames.

Table 4.1 summarizes the methods for facial expressions classification, the number of subjects used to make the evaluation and their performances. It is very difficult to compare between the methods described above. They are based on different modeling of the temporal information and they do not achieve the same type of temporal classification.

Cohen *et al* aim at separating between a sequence of different expressions; Busso *et al* recognize a sequence of facial expression by the combination of the static classification on each frame of the sequence; Zhang *et al* classify the sequence frame by frame by using at each time the last classification results; and finally Pantic *et al* classify temporal segments of

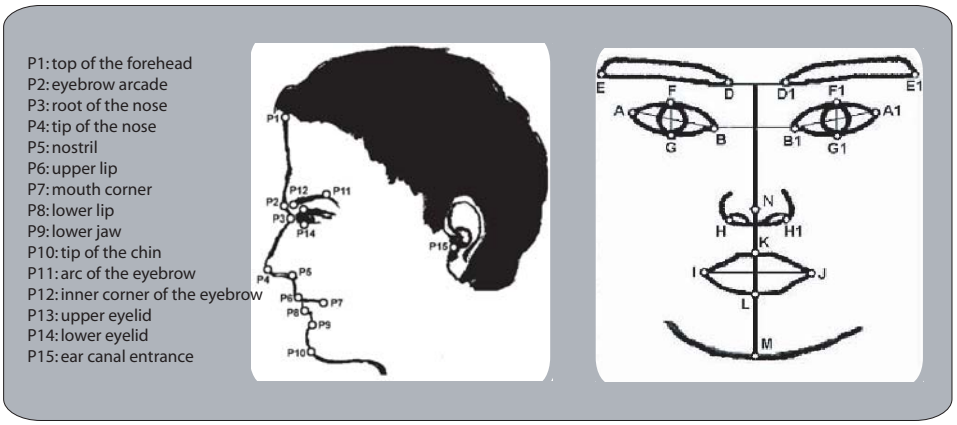

Figure 4.4: Left: profile facial points [Pan06]; right: frontal facial points [Pan05a]

Action Units. However, the common limitation of these methods is that they are based on the integration or the combination of results based on static classification or, at best, only on the information provided by the last frame is taken into account in the classification process. The dynamic evolution of the facial features states during a sequence of facial expression is not taken into account.

Reference	Classification	Nbr of AUs or expressions	Nbr of test subjects	Performances (%)
[Coh03b]	HMMs	6	5	55.46
[Bus04]	K-nearest neighbor (KNN)	4	1	85
[Zha05]	Dynamic Bayesian network (DBN)	6	–	–
[Pan05a]	Rule-based method	27 AUs and combination	MMI [Pan05b] + Cohn-Kanade databases	Based on: Profile view 93.6 Frontal view 90

Table 4.1: Comparisons of dynamic facial expression recognition algorithms.

In the next section we present our preliminary work on the introduction of the temporal information for the classification of facial expressions consisting in, firstly, combining at each time t the current information with the one obtained at time $t-1$ and secondly, in recognizing sequence of facial expressions as a sequence of facial features deformations between two *Neutral* states.

4.1 Temporal information for facial expressions classification

A fist step in the introduction of the temporal information is to take into account at each time t the available information at time t-1. Indeed the information between two consecutive frames is strongly related because an expression is the result of a progressive evolution of the facial features deformations. Moreover based on automatic segmented data, the introduction of a temporal constraint allows to reduce punctual errors occurring during the evolution of the characteristic distances states.

Indeed distance value can be much higher in absolute than its real value. In order to reduce the importance of these high jumps, first, a smoothing process (Gaussian smoothing) is applied on the values of the characteristic distances states. However even if it manages to reduce the intensity of the errors, it cannot delete them totally, especially when the distance evolution is in the opposite of its real evolution (C^- instead of C^+). A temporal constraint on the characteristic distances states would also allow to overcome these kind of errors.

For all these reasons, temporal information is introduced in the form of conditional pieces of evidence, gathered in a transition matrix, for each characteristic distance to be applied in a sequence of facial expressions.

4.1.1 Conditional pieces of evidence

We introduce the definition of conditional pieces of evidence so as to apply a temporal constraint on the computation of the basic belief assignment associated to each characteristic distance. The conditional piece of evidence corresponds to the pieces of transitions from each proposition k of the frame of discernment at time $t-1$ (previous frame) to each one of the possible propositions r at time t (current frame) and can be noted $m(r/k)$ (k and $r \in \{S$, C^+,C^-,$S \cup C^+$,$S \cup C^-)$. For example, for a considered distance D_j, $m_{D_j}(C^+/S)$ corresponds to the piece of evidence (the belief) $m_{D_j}(C^+)$ at time t if $m_{D_j}(S) = 1$ at time $t-1$.

For each characteristic distance D_j ($1 \leq j \leq 5$), all the conditional pieces of evidence are gathered in a matrix called the *transition matrix*, noted $\overline{M}$:

$$\overline{M}(D_j) = \begin{pmatrix} m(S/S) & m(S/C^+) & m(S/C^-) & m(S/S \cup C^+) & m(S/S \cup C^-) \\ m(C^+/S) & m(C^+/C^+) & m(C^+/C^-) & m(C^+/S \cup C^+) & m(C^+/S \cup C^-) \\ m(C^-/S) & m(C^-/C^+) & m(C^-/C^-) & m(C^-/S \cup C^+) & m(C^-/S \cup C^-) \\ m(S \cup C^+/S) & m(S \cup C^+/C^+) & m(S \cup C^+/C^-) & m(S \cup C^+/S \cup C^+) & m(S \cup C^+/S \cup C^-) \\ m(S \cup C^-/S) & m(S \cup C^-/C^+) & m(S \cup C^-/C^-) & m(S \cup C^-/S \cup C^+) & m(S \cup C^-/S \cup C^-) \end{pmatrix} \tag{4.1}$$

where the sum of all the conditional pieces of evidence belonging to the same column is equal to 1 and the matrix dimensions ([row X column]) is $[5X5]$.

Our aim is to obtain the transition matrix $\overline{M}(D_j)$ associated to each characteristic distance and which is independent of the subject and of the expression.

The conditional pieces of evidence are computed on the results of the manual segmentation of the HCE training database (11 subjects and 3 expressions). For each video sequence we have a basic belief assignment m_k associated to each frame k as:

$$m_k = \begin{pmatrix} m(S)_k \\ m(C^+)_k \\ m(C^-)_k \\ m(S \cup C^+)_k \\ m(S \cup C^-)_k \end{pmatrix} \tag{4.2}$$

Then a conditional basic belief assignment can be defined between each two consecutive frames: between the first and second frames ($m_1 \rightarrow m_2$), between the second and third frames ($m_2 \rightarrow m_3$) and generalizing between the kth and $k+1$th frames ($m_k \rightarrow m_{k+1}$, $1 \leq k \leq 120$, 120 is the number of frames per sequence).

Then it exists a transition matrix for each distance D_j, for each subject i and each expression e noted $\overline{M}_i^e(D_j)$ such as:

$$m_{k+1} = \overline{M}_i^e(D_j) m_k \tag{4.3}$$

Equation 4.3 is defined for one transition. To obtain the transitions on the whole of the sequence we concatenate the m_k as:

$$\begin{pmatrix} m(S)_2 & \dots & m(S)_{120} \\ m(C^+)_2 & \dots & m(C^+)_{120} \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ m(S \cup C^-)_2 & \dots & m(S \cup C^-)_{120} \end{pmatrix}_{[5X119]}^{e}_{i} = \overline{M}_i^e(D_j) * \begin{pmatrix} m(S)_1 & \dots & m(S)_{119} \\ m(C^+)_1 & \dots & m(C^+)_{119} \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ m(S \cup C^-)_1 & \dots & m(S \cup C^-)_{119} \end{pmatrix}_{[5X119]}^{e}_{i} \tag{4.4}$$

$$M^e_{i,(2..120)}(D_j) = \overline{M}_i^e(D_j) * M^e_{i,(1..119)}(D_j) \tag{4.5}$$

where:
$M^e_{i,(2..120)}(D_j)$ corresponds to the basic belief assignment of the considered distance D_j, for the subject i from the frame 2 until 120 for the expression e.
$M^e_{i,(1..119)}(D_j)$ corresponds to the basic belief assignment of the considered distance D_j, for the same subject i from the frame 1 until 119 for the same expression e.
The elements of each column of these matrices correspond to basic belief assignment associated to the characteristic distance (D_j) for the considered subject i and the expression e in the current frame. For example for the proposition S in $M^e_{i,(2..120)}(D_j)$, $m(S)_2$ corresponds to the piece of evidence of the state S for the subject i in the frame 2.

Then for each distance we obtain a transition matrix associated to each subject and to each expression. However our aim is to define a transition matrix (see equation 4.1) for each distance independently of the subjects and independently of the studied expressions. To do this we concatenate the matrices $M^e_{i,(2..120)}(D_j)$ (resp. $M^e_{i,(1..119)}(D_j)$) of all the subjects i ($1 \leq i \leq 11$) and of all the expressions e ($e \in \{Smile, Surprise, Disgust\}$) to define the transition matrix $\overline{M}(D_j)$ ($1 \leq j \leq 5$) for each distance D_j as:

$$\underset{[5X(119*3*11)]}{\left(M^{sm}_{1,(2..120)} \quad M^{sp}_{1,(2..120)} \quad M^{dg}_{1,(2..120)} \quad ... \quad M^{sm}_{11,(2..120)} \quad M^{sp}_{11,(2..120)} \quad M^{dg}_{11,(2..120)}\right)} =$$

$$\underset{[5X5]}{\overline{M}(D_j)} * \underset{[5X(119*3*11)]}{\left(M^{sm}_{1,(1..119)} \quad M^{sp}_{1,(1..119)} \quad M^{dg}_{1,(1..119)} \quad ... \quad M^{sm}_{11,(1..119)} \quad M^{sp}_{11,(1..119)} \quad M^{dg}_{11,(1..119)}\right)}$$

$$\begin{aligned} M^{sm,sp,dg}_{(1..11),(2..120)} &= \overline{M}(D_j) * M^{sm,sp,dg}_{(1..11),(1..119)} \\ \overline{M}(D_j) &= M^{sm,sp,dg}_{(1..11),(2..120)} * (M^{sm,sp,dg}_{(1..11),(1..119)})^{-1} \end{aligned} \tag{4.6}$$

where sm: *Smile*, sp:*Surprise* and dg:*Disgust*.

Equation 4.6 leads to an overdetermined system where $\overline{M}(D_j)$ is its optimum solution in the least square sense.

It has to be noted that in the training database according to the subject, each expression is expressed with different evolution speed (for example we have slow *Smile* as well as quick *Smile*). Then the obtained transition matrix takes into account fast as well as slow evolution of the facial features deformations in a facial expression sequence.

4.1.2 Discounting

In the sequences of the HCE database, the subjects were asked to simulate an expression during a specific temporal window (5s). As a consequence it appears that they tended to remain artificially longer in an expressive state which is traduced by the fact that characteristic distances remains longer in a singleton state than in doubt states. Then the number of transitions from each state to itself is much higher than its transition to the other states. This leads to overweighted transitions for the diagonal elements of $\overline{M}(D_j)$ $(1 \leq i \leq 11)$(for example $m(C^+/C^+)$ is more than 90%). To avoid this prediction bias due to the simulation, a *discounting* is applied to the transition matrix. It allows to remove incorrect transitions while preventing the overweighting of other transitions. The *discounting* [Sme00] consists in defining a parameter $\alpha \in [0, 1]$ which allows to compute the new piece of evidence $m^{'}$ of each proposition A_i according to its current piece of evidence as:

$$\widetilde{m}(A_i) = \alpha m(A_i) \tag{4.7}$$

$$\widetilde{m}(A_i \cup \overline{A_i}) = 1 - \alpha(1 - m(A_i \cup \overline{A_i})) \tag{4.8}$$

Where $\overline{A_i}$ corresponds to the complement of A_i.

In the scope of our application A_i corresponds to the elements of the transition matrix $\overline{M}(D_j)$ (see Equation 4.1). A modified version of the discounting is then used to discount the diagonal conditional pieces of evidence of the transition matrix and to redistribute their subtracted values on the doubt states. The other transition states do not change. For example

for the first row of $\overline{M}(D_j)$ the discounting consists in computing:

$$\begin{aligned}
\widetilde{m}(S/S) &= \alpha m(S/S) \\
\widetilde{m}(S \cup C^+/S) &= 1 - \alpha(1 - m(S \cup C^+/S)) \\
\widetilde{m}(S \cup C^-/S) &= 1 - \alpha(1 - m(S \cup C^-/S)) \\
\widetilde{m}(C^+/S) &= m(C^+/S) \\
\widetilde{m}(C^-/S) &= m(C^-/S)
\end{aligned}$$

$\widetilde{m}$ is a basic belief assignment then the sum of all the pieces of evidence has to be equal to 1. If it is not the case we redistribute the difference equally on the pieces of evidence of the doubt states $\widetilde{m}(S \cup C^+/S)$ and $\widetilde{m}(S \cup C^-/S)$.

The value of α needs to be sufficiently important to discount the diagonal states of $\overline{M}(D_j)$ and sufficiently weak not to change the basic belief assignment of the conditional pieces of evidences of the doubt states. In our case we choose $\alpha = 0.8$.

Based on the equation 4.6 and after discounting, the computation leads to the following transition matrices associated to the five characteristic distances:

$$\overline{M}(D_1) = \begin{pmatrix}
\mathbf{0.60} & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & \mathbf{0.57} & 0.00 & \mathbf{0.26} & 0.00 \\
0.00 & 0.00 & \mathbf{0.61} & 0.00 & \mathbf{0.11} \\
\mathbf{0.11} & \mathbf{0.43} & 0.00 & \mathbf{0.74} & 0.00 \\
\mathbf{0.29} & 0.00 & \mathbf{0.39} & 0.00 & \mathbf{0.89}
\end{pmatrix}$$

$$\overline{M}(D_2) = \begin{pmatrix}
\mathbf{0.61} & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & \mathbf{0.60} & 0.00 & \mathbf{0.21} & 0.00 \\
0.00 & 0.00 & \mathbf{0.61} & 0.00 & \mathbf{0.11} \\
\mathbf{0.14} & \mathbf{0.40} & 0.00 & \mathbf{0.79} & 0.00 \\
\mathbf{0.25} & 0.00 & \mathbf{0.39} & 0.00 & \mathbf{0.89}
\end{pmatrix}$$

$$\overline{M}(D_3) = \begin{pmatrix}
\mathbf{0.61} & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & \mathbf{0.60} & 0.00 & \mathbf{0.19} & 0.00 \\
0.00 & 0.00 & \mathbf{0.60} & 0.00 & \mathbf{0.12} \\
\mathbf{0.22} & \mathbf{0.40} & 0.00 & \mathbf{0.81} & 0.00 \\
\mathbf{0.17} & 0.00 & \mathbf{0.40} & 0.00 & \mathbf{0.88}
\end{pmatrix}$$

$$\overline{M}(D_4) = \begin{pmatrix}
\mathbf{0.60} & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & \mathbf{0.60} & 0.00 & \mathbf{0.22} & 0.00 \\
0.00 & 0.00 & \mathbf{0.53} & 0.00 & \mathbf{0.10} \\
\mathbf{0.30} & \mathbf{0.40} & 0.00 & \mathbf{0.78} & 0.00 \\
\mathbf{0.10} & 0.00 & \mathbf{0.47} & 0.00 & \mathbf{0.90}
\end{pmatrix}$$

$$\overline{M}(D_5) = \begin{pmatrix}
\mathbf{0.60} & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & \mathbf{0.59} & 0.00 & \mathbf{0.15} & 0.00 \\
0.00 & 0.00 & \mathbf{0.61} & 0.00 & \mathbf{0.12} \\
\mathbf{0.18} & \mathbf{0.41} & 0.00 & \mathbf{0.85} & 0.00 \\
\mathbf{0.22} & 0.00 & \mathbf{0.39} & 0.00 & \mathbf{0.88}
\end{pmatrix}$$

4.1.3 Predicted basic belief assignment

The predicted basic belief assignment consists in predicting the pieces of evidence of the characteristic distances states $m_{pred,t}(D_j)$ $(1 \leq j \leq 5)$ at time t according to their pieces of evidence at time $t-1$. The predicted (*pred*) basic belief assignment is then computed at time t by the combination of the transition matrix $\overline{M}(D_j)$ and the basic belief assignment computed (*comp*) at time $t-1$ in the following way:

$$m_{pred,t}(D_j) = \overline{M}(D_j) m_{comp,t-1}(D_j) \tag{4.9}$$

Where:

$$m_{pred,t}(D_j) = \begin{pmatrix} m(C^+) \\ m(C^-) \\ m(S) \\ m(S \cup C^+) \\ m(S \cup C^-) \end{pmatrix}_t \quad m_{comp,t-1}(D_j) = \begin{pmatrix} m(C^+) \\ m(C^-) \\ m(S) \\ m(S \cup C^+) \\ m(S \cup C^-) \end{pmatrix}_{t-1} \tag{4.10}$$

$m_{pred,t}(D_j)$ is the predicted basic belief assignment of the characteristic distances states at time t and $m_{comp,t-1}(D_j)$ is the computed basic belief assignment at time $t-1$.

4.1.4 Combination of the computed and predicted pieces of evidence

At time t, the dynamic correction of the pieces of evidence associated with each characteristic distance state combines two processes: computation of the BBA of the characteristic distances states at time t ($m_{comp,t}$) and their temporal prediction ($m_{pred,t}$) based on the previous frame at time $t-1$ using the transition matrix (see Equation 4.9). The combination of $m_{comp,t}$ and $m_{pred,t}$ is based on the *conjunctive combination* (orthogonal sum, see Equation 3.11) as:

$$m_{comb,t} = m_{pred,t} \oplus m_{comp,t} \tag{4.11}$$

Then at time t Table 4.2 is applied to define the new pieces of evidence according to the predicted (*pred*) and the computed (*comp*) pieces of evidence (see also section 3.7.3.3).

Pred \ Comp	S	C^+	C^-	$S \cup C^+$	$S \cup C^-$
S	S	ϕ	ϕ	S	S
C^+	ϕ	C^+	ϕ	C^+	ϕ
C^-	ϕ	ϕ	C^-	ϕ	C^-
$S \cup C^+$	S	C^+	ϕ	$S \cup C^+$	S
$S \cup C^-$	S	ϕ	C^-	S	$S \cup C^-$

Table 4.2: Combination between the pieces of evidence of the predicted and of the estimated states. ϕ denotes a conflict state.

The combination leads sometimes to a conflict, noted ϕ, between the predicted and the computed pieces of evidence ($m_{pred,comp}(\phi) \neq 0$). This is mainly due to segmentation errors so in this case, the results obtained by prediction $m_{pred,t}$ is chosen to form the basic belief assignment associated to the distances states at time t.

In the following the proposed classification methods are based on the data after combination between the computed and predicted pieces of evidence.

4.2 Frame-by-Frame expressions classification

The classification is based on the basic belief assignment of the characteristic distances states after the combination of the predicted and the computed pieces of evidence (see section 4.1.4). Then the classification is made similarly to the static classification (see section 3.7.3.3). This classification method is called **Frame-by-Frame expressions classification**.

Figure 4.5 presents an example of the Frame-by-Frame classification results on a *Smile* sequence. We can see the sensitivity of the proposed method to the evolution of the facial features behaviors under different intensities. In each frame we have two plots: the current image on the left and the corresponding expressions and their pieces of evidence on the right. The first plot of the figure corresponds to the second frame of the sequence. It is recognized as *Neutral* expression at 100%. The second plot corresponds (frame 15 of the sequence) to a transition state when the subject is neither in the *Neutral* state, neither in the studied expression. The system classifies it as 77% *Neutral* and 23% *Unknown*. In the third plot (frame 17 of the sequence) the system starts recognizing the expression of *Smile*. Finally the last plot (frame 25 of the sequence) corresponds to the apex of *Smile* expression.

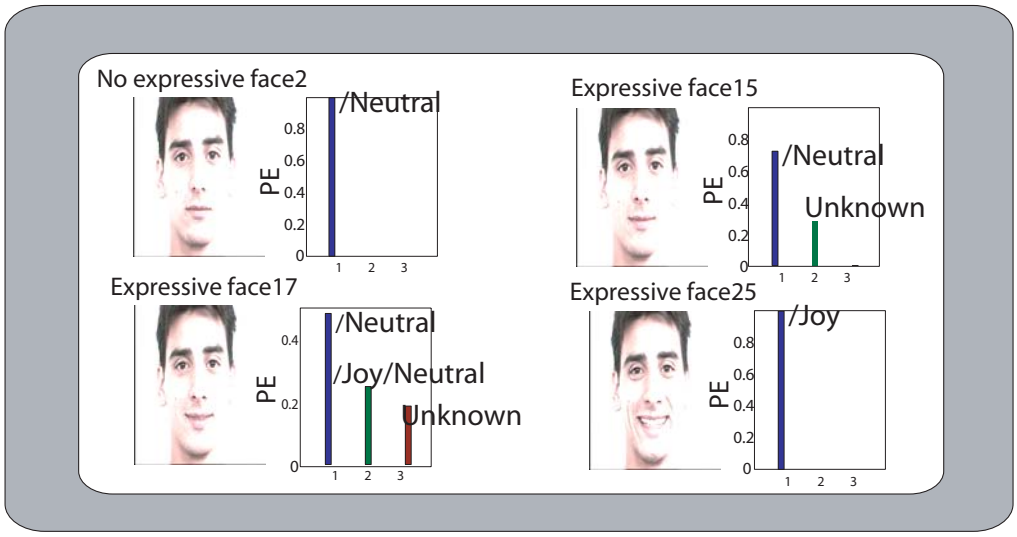

Figure 4.5: Results of the Frame-by-Frame expressions classification. In each frame, left: current frame; right: all the facial expressions with a not null piece of evidence.

In the presented Frame-by-Frame classification, the temporal information is only limited to time $t-1$. In the following we aim at obtaining a dynamic classification of the whole facial expression sequence from the beginning until the end of the facial features deformations.

4.3 Dynamic expressions classification

A facial expression is the result of progressive deformations of a set of facial features which can appear at different time and without any defined appearance order (asynchronously) (see Figure 4.7 for *Surprise* expression). The proposed method allows to deal with these considerations by the dynamic analysis of the characteristic distances states. Each facial expression is characterized by a *beginning*, an *apex* (it corresponds to the maximum of intensity

reached by the facial expression associated with a particular distances states configuration) and an *end* (see Figure 4.6). Here we present a method called **Dynamic expressions classification** to recognize a sequence of facial expressions between each pair of *beginning* and *end*.

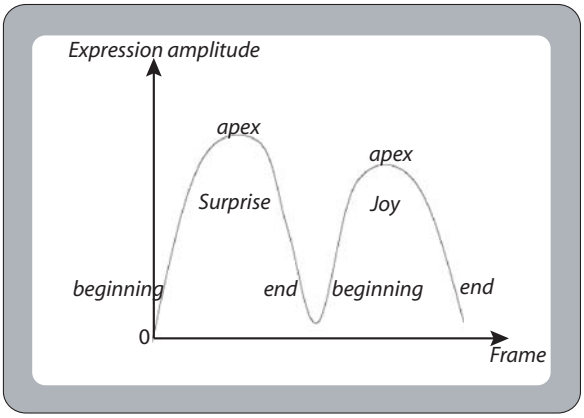

Figure 4.6: Example of sequence displaying two expressions sequences, *Surprise* and *Smile*. "0" stands for *Neutral* expression.

In each expression sequence, the *beginning* is detected as the first frame where at least one of the characteristic distances is no more in the stable state S; the *end* is detected as the first frame where all the states distances have come back to the stable state S. There is no means to identify the *apex* of the expression because the expressions are identified by a combination of a characteristic distances states between each *beginning* and *end*.

Once the *beginning* of the expression detected, the analysis of the distances states is made inside an increasing temporal window. Figure 4.7 shows an example of evolution of the increasing temporal window during the analysis of *Surprise* expression sequence. The size of the window increases progressively at each time between the *beginning* and the *end* of the expression. Then, at each time the whole set of the previous information (the past states of the characteristic distances) is taken into account to classify the current expression sequence. Once the *beginning* is detected the classification consists in defining at each time the basic belief assignment of the characteristic distances states according to their past basic belief assignments from the *beginning* until the current frame. To do this, at each time t, inside the current increasing window and for each characteristic distance, a criteria has to be used to select its corresponding state. The selection is made according to two parameters: the number of appearance of each symbolic state $K \in \{C^+, C^-, S \cup C^+, S \cup C^-\}$ noted $\underset{\Delta t}{Nb}(K)$ and their integral (sum) of plausibility noted $\underset{\Delta t}{PL}(K)$ (see Equation 3.13) computed inside the temporal increasing window (Δt).

For example for $state = C^+$:

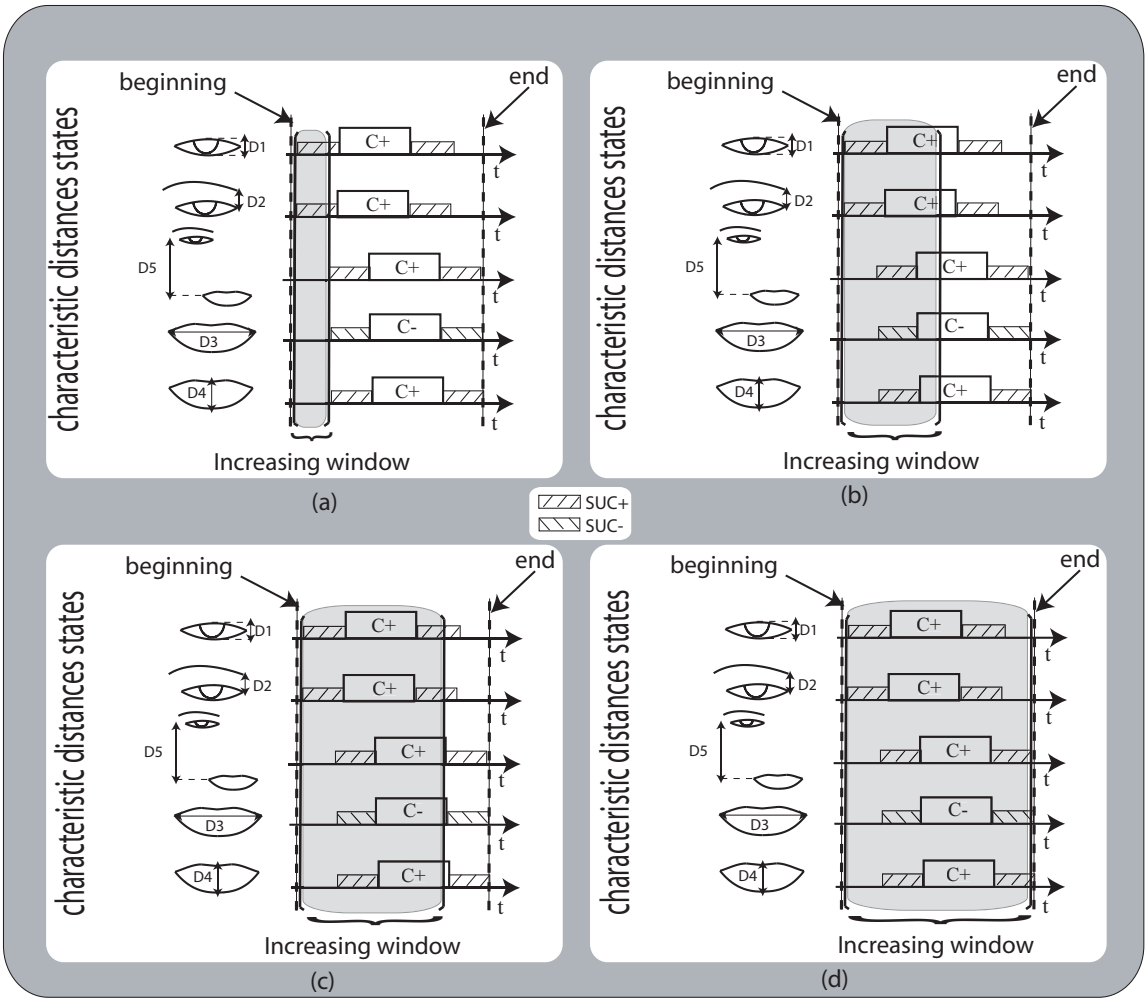

Figure 4.7: Example of the increasing temporal window during a sequence of *Smile* expression.

$$K_i(C^+) = \begin{cases} 1 & \text{if } m(C^+) \neq 0 \\ 0 & \text{otherwise} \end{cases}, 1 \leq i \leq \Delta t$$

$$\underset{\Delta t}{Nb}(C^+) = \sum_{i=1}^{\Delta t}(K_i(C^+))$$

$$\underset{\Delta t}{PL}(C^+) = \sum_{\Delta t}(m(C^+) + m(S \cup C^+))$$

with $K_i \in \{C^+, C^-, S \cup C^+, S \cup C^-\}$ the symbolic state of frame i

The study of the whole set of the distances states configurations on the HCE_T database allows us to define the rules Table 4.3. It summarizes the rules based on the values of $\underset{\Delta t}{Nb}(K)$ and $\underset{\Delta t}{PL}(K)$ used to choose the distances states at each time inside the temporal increasing

	$\underset{\Delta t}{Nb}(C^+)$	$\underset{\Delta t}{Nb}(C^-)$	$\underset{\Delta t}{Nb}(S \cup C^+)$	$\underset{\Delta t}{Nb}(S \cup C^-)$	$\underset{\Delta t}{PL}(C^+) - \underset{\Delta t}{PL}(C^-)$	$\underset{\Delta t}{PL}(S \cup C^+) - \underset{\Delta t}{PL}(S \cup C^-)$
S	$= 0$	$= 0$	$= 0$	$= 0$	/	/
C^+	$\neq 0$	$= 0$	/	/	/	/
	$\neq 0$	$\neq 0$	/	/	> 0	/
C^-	$= 0$	$\neq 0$	/	/	/	/
	$\neq 0$	$\neq 0$	/	/	< 0	/
$S \cup C^+$	$= 0$	$= 0$	$\neq 0$	$= 0$	/	/
	$= 0$	$= 0$	$\neq 0$	$\neq 0$	/	> 0
$S \cup C^-$	$= 0$	$= 0$	$= 0$	$\neq 0$	/	/
	$= 0$	$= 0$	$\neq 0$	$\neq 0$	/	< 0

Table 4.3: Rules table for the chosen states inside a sliding window Δ_t (/: not used). Rows correspond to the chosen propositions in the sliding window; columns correspond to the required conditions.

window. The rows correspond to the rules associated to each distance state and the columns to the states of the required conditions obtained inside the temporal increasing window. The rules table is defined according to three rules:

- If only one singleton state appears inside the increasing window, this one is chosen to be the state of the studied characteristic distance. Figure 4.8 (a) shows an example, in this case the state C^+ is chosen,
- If two singleton states appear, the most plausible state between them is chosen. Figure 4.8 (b) shows an example, in this case the state C^- is chosen,
- If only doubt states appear, the most plausible one between them is chosen. Figure 4.8 (c) shows an example, in this case the state $S \cup C^-$ is chosen.

At the beginning all the distances are in the state S and change only if one of the other states appear in the increasing window. In this case the corresponding state is chosen according to the rules defined in Table 4.3. Two rows are associated to each proposition which is chosen if the conditions (columns) corresponding to one of these two rows are checked. For example C^- is chosen if:

- the number of occurrence of C^- in the increasing window is different from zero ($\underset{\Delta t}{Nb}(C^-) \neq 0$) and the number of occurrence of C^+ is equal to zero ($\underset{\Delta t}{Nb}(C^+) = 0$).

or if

- the integral of plausibility of C^- is higher than the one of C^+: $\underset{\Delta t}{PL}(C^+) - \underset{\Delta t}{PL}(C^-) < 0$.

The piece of evidence associated to each chosen state corresponds to its maximum value inside the current temporal increasing window. Finally at time t (between the *beginning* and the *end* of the expression sequence), once the basic belief assignment of all the characteristic distances are computed, the classification is carried out based on the same rules table (see Figure 3.26). To summarize, a decision can be made at each time t taking into account

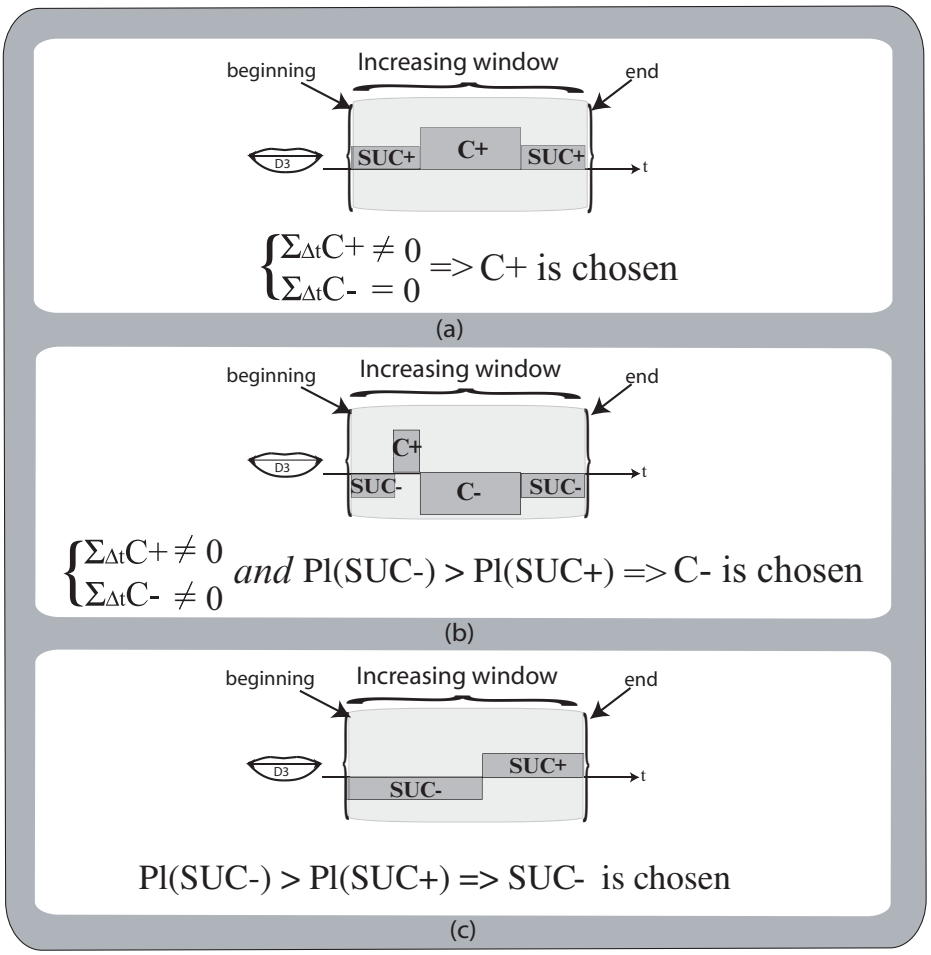

Figure 4.8: Selection process of the characteristic distances states inside the increasing temporal window .

the past basic belief assignment of the characteristic distances states from the *beginning* to the current frame. Then the end of the expression (the current frame is the last frame of the sequence), the decision is made taking into account all the past basic belief assignments of the characteristic distances states and gives the classification of the entire expression sequence.

4.4 Experimental results

In this section we present the results of our preliminary works on the introduction of temporal information in the classification of the facial expressions. First we compare the results of the proposed Frame-by-Frame classification with the results obtained on our previous works without the use of the transition matrix (i.e static information). Then we present the results of the Dynamic classification. All the experiments have been realized on the HCE_T on data obtained with our segmentation algorithms.

4.4.1 Facial features deformations analysis

So as to summarize all the information extracted and analyzed during an expression sequence, we describe here a complete interface.

Since the proposed approach deals with the face expression evolution as well as the description of the permanent facial features deformations (eyes, eyebrows and mouth), the interface first displays an analysis of the individual facial features deformations. Notably this analysis gives a description of the eyes, the eyebrows and the mouth state with their piece of evidence even in case of *Unknown* expressions. It has to be stressed that this description is particularly interested in case of *Unknown* expression, because even in that case it gives a description of the facial features states. To do this, each characteristic distance state is translated into deformations of the corresponding facial features (with its corresponding piece of evidence). Only the singleton states (i.e S, C^+ or C^-) are considered to define the facial features deformations. In the case of doubt states we prefer not to take a decision. This translation is defined by the following set of rules:

- **Eyes analysis:** Slackened eyes ($D_1 == S$), Open eyes ($D_1 == C^+$), Eyes slightly closed ($D_1 == C^-$),

- **Eyebrows analysis:** Slackened eyebrows ($D_2 == S$), Raised inner eyebrows ($D_2 == C^+$), Lowered inner eyebrows ($D_2 == C^-$),

- **Mouth analysis:** Closed mouth ($D_4 == S$), Open mouth ($D_4 == C^+$), Mouth corners pulled downwards ($D_5 == C^+$), Mouth corners pulled up ($D_5 == C^-$).

The characteristic distance D_3 is not used because sometimes it can increase while the mouth remains closed.

Figure 4.9 presents an example of the information displayed during the analysis of a facial expression sequence and the corresponding facial features deformations. This example shows the results on the frame 44 from a *Smile* sequence composed of 120 frames. It corresponds to the apex of the expression. The interface is divided into five different regions: on top left, the current frame to be analyzed; on top middle the result of the Frame-by-Frame classification (here this is a *Smile* expression with a piece of evidence equal to 100%); on top right, the result of the dynamic classification which corresponds to the classification of the sequence since the beginning until the current frame (here *Smile* sequence); on bottom left, the current states of the characteristic distances and their pieces of evidence; on bottom right, the corresponding facial features deformations (for example *slackened eyebrows* corresponds to D_2 in state S with its corresponding piece of evidence 100%). No information on the eye state is reported because the corresponding characteristic distance is in a doubt state ($S \cup C^-$).

4.4.2 Frame-by-Frame classification performance

The aim of this section is to discuss the improvement of the Frame-by-Frame classification induced by the introduction of the temporal information (conditional pieces of evidence) and to compare it with our previous static classification on data obtained from our automatic segmentation (see Table 4.4).

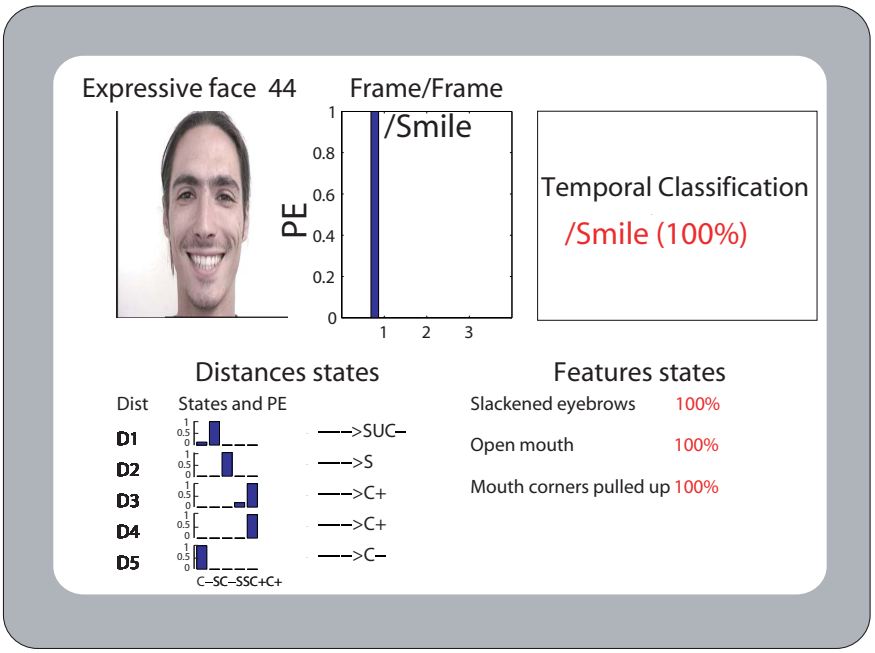

Figure 4.9: Classification result interface displaying: distances states estimation; Frame-by-Frame classification; facial features deformations and dynamic classification with the associated pieces of evidence on frame 44 being part of a *Smile* sequence.

The analysis of the Frame-by-Frame classification on the facial expression sequences shows that the mean improvement of the results according to our previous results are 9%.

Smile expression recognition rate increases of 13.6%. It is interesting to observe that doubt expression *Smile-Disgust* decreases in favor of singleton *Smile* expression. In this case the temporal information allows to improve the reliability of the results compared with those of the static classification decreasing the doubt between expressions.

In case of *Surprise*, we have observed in the static classification that this expression was often misclassified in favor of *Fear* expression. Here *Fear* decreases in favor of *Surprise-Fear* which increases of 14%. Then the introduction of the temporal information allows to reduce these errors in favor of doubt which is actually satisfactory because as explained before these two expressions are difficult to distinguish.

Disgust recognition rate is not modified by the temporal information. As explained before the low classification rates of this expression are not due to segmentation errors (they are comparable to the rates obtained on manual segmentation, see Table 3.9) but to the difficulty of simulating this expression by non-actor people.

Figure 4.10 shows examples of Frame-by-Frame classification results on *Surprise*, *Smile* and *Disgust* expression sequences. Two frames are presented for each sequence. The first row presents the facial features analysis and their corresponding expression for the first frame and the apex of *Disgust* expression sequence. The last two rows show the same information for the intermediate state and the apex for *Smile* and *Disgust* expression sequences respectively. We can observe that the facial features states are reported only where the corresponding distances

System \ Expert	E_1	E_2	E_3	E_7
E_1 *Smile*	29.6	0	0	0
E_2 *Surprise*	0	0	0	0
E_3 *Disgust*	4.8	0	43	0
E_4 *Anger*	0	0	3	0
E_5 *Sadness*	1	0	0	1.2
E_6 *Fear*	0	5.2	0	0
E_7 *Neutral*	2.70	1.1	2.10	72
E_8 *Unknown*	7.50	4.70	38	5
$E_1 \cup E_3$	42.4	9	4.5	4
$E_2 \cup E_6$	0	75.4	0	0
others	12	9	9.2	17.8
Total	**72**	**75.4**	**47.5**	**72**

Table 4.4: Classification rates in percent with data obtained from automatic segmentation for the HCE_T database.

are in a singleton state (the analysis is not reported in the case of doubt states).

The analysis of the Frame-by-Frame classification results shows that the temporal information improves the classification results by correcting segmentation errors but not all the errors. Most of the time these remaining errors are due to the detection of the mouth corners leading to errors in the state of the characteristic distance D_3 which explains the augmentation of the rates of the row *Others* (see Table 4.4) in comparison with the manual data. Indeed mouth corners are difficult to segment robustly [Eve04]. One solution to overcome this problem may be to discount D_3 during the fusion of all the characteristic distance states so as to take into account the lack of reliability associated to this distance.

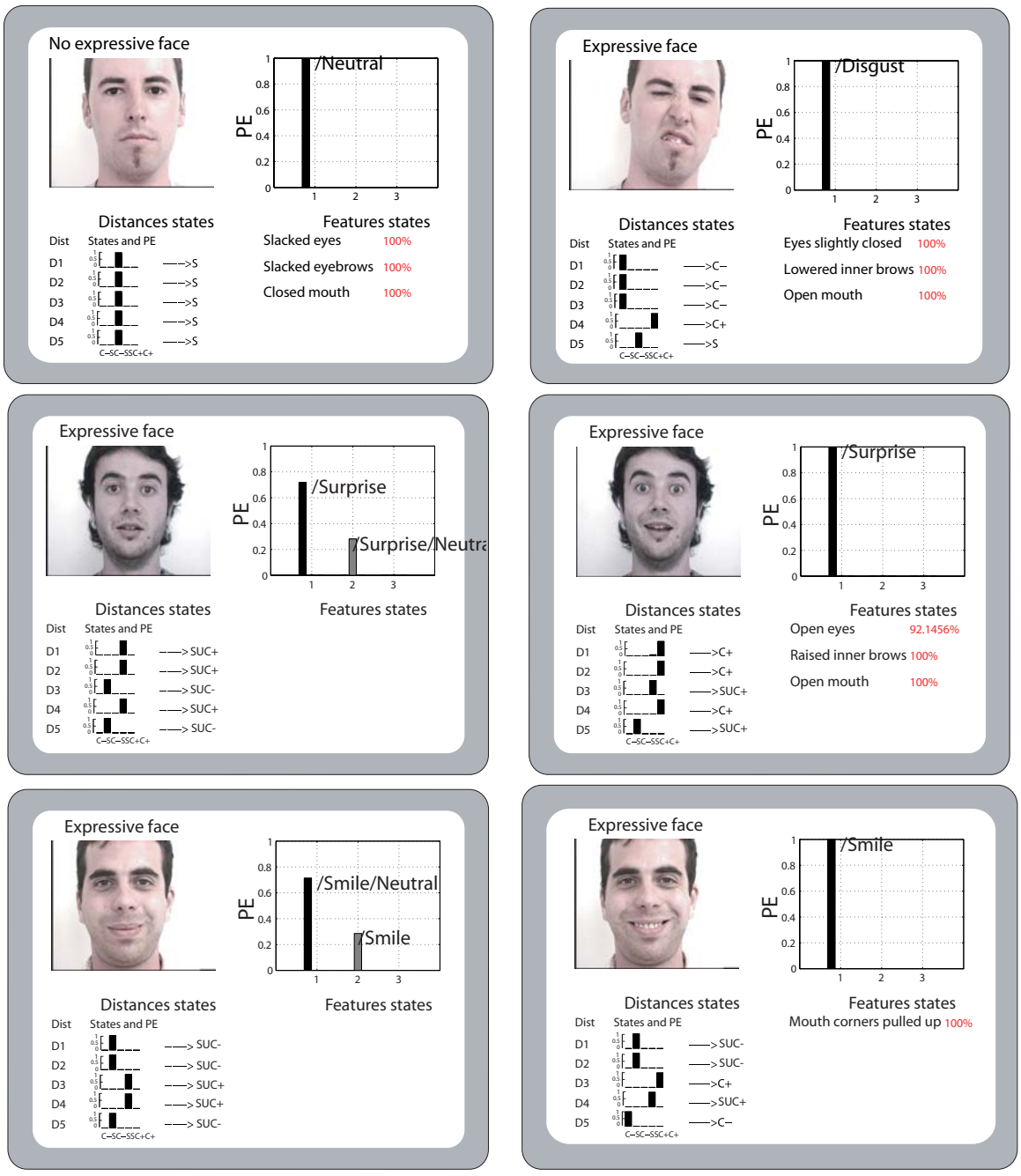

Figure 4.10: Examples of Frame-by-Frame classification results and their corresponding facial features analysis for *Smile*, *Surprise* and *Disgust* expressions.

4.4.3 Dynamic classification performances

The dynamic classification has been tested on the HCE_T database (24 sequences, 8 sequences for each of 3 expressions). In order to evaluate the robustness of the proposed method we have tested our algorithm on the manual as well as on the automatic segmentation results. Classification rates on the three expressions and for the two types of data are given in Table 4.5.

We can observe only the expressions and the associated doubt state appear for the two types of data. All the other states never appear in the final classification showing the robustness of the dynamic classification. Moreover, the classification rates are very comparable.

Automatic segmentation			
System \ Expert	E_1	E_2	E_3
E_1 *Smile*	37.5	0	0
E_2 *Surprise*	0	75	0
E_3 *Disgust*	0	0	25
E_4 *Anger*	0	0	0
E_5 *Sadness*	0	0	0
E_6 *Fear*	0	0	0
E_7 *Neutral*	0	0	0
E_8 *Unknown*	0	25	25
$E_1 \cup E_3$	62.5	0	50
$E_2 \cup E_6$	0	0	0
others	0	0	0
Total	**100**	**75**	**75**

Manual segmentation			
System \ Expert	E_1	E_2	E_3
E_1 *Smile*	37.5	0	0
E_2 *Surprise*	0	88	0
E_3 *Disgust*	0	0	50
E_4 *Anger*	0	0	0
E_5 *Sadness*	0	0	0
E_6 *Fear*	0	0	0
E_7 *Neutral*	0	0	0
E_8 *Unknown*	0	12	25
$E_1 \cup E_3$	62.5	0	25
$E_2 \cup E_6$	0	0	0
others	0	0	0
Total	**100**	**88**	**75**

Table 4.5: Dynamic classification rates in percent based on: left, results on automatic segmentation, right, results on manual segmentation.

Figure 4.11 presents an example of a dynamic classification on a *Surprise* sequence. The original sequence has 120 frames where the subject evolves from *Neutral*, reaches *Surprise* and comes back to *Neutral*. We give selected frames to convey our results which shows the evolution over time of the facial features deformations.
In frame 1 the subject is in the *Neutral* state. At this time the system cannot give any temporal classification results. In frame 13 we can observe the sensitivity of the system to recover the behavior of the facial features. Based on the states of the characteristic distances, the Frame-by-Frame classification confidence is 80% which corresponds to a *Surprise* expression and 20%, to a doubt between *Surprise* and *Neutral*. The temporal classification corresponds to the classification result on all the frames from the beginning until the current frame 13. At this time, the temporal classification confidence is 81% which corresponds to a *Surprise* sequence. In frame 26, the Frame-by-Frame result as well as the temporal classification reaches a confidence level of 100% on the recognition of a *Surprise* expression.

The last two frames 65 and 76 give the classification results when the subject comes back to the *Neutral* state. We can observe the evolution of facial features deformations coming back to slackened states and the Frame-by-Frame classification giving a *Neutral* state on the last frame. However we can notice that the temporal classification does not change and gives

the classification over the whole sequence.

These rates correspond to preliminary results. Tests on databases containing more examples are required to confirm them.

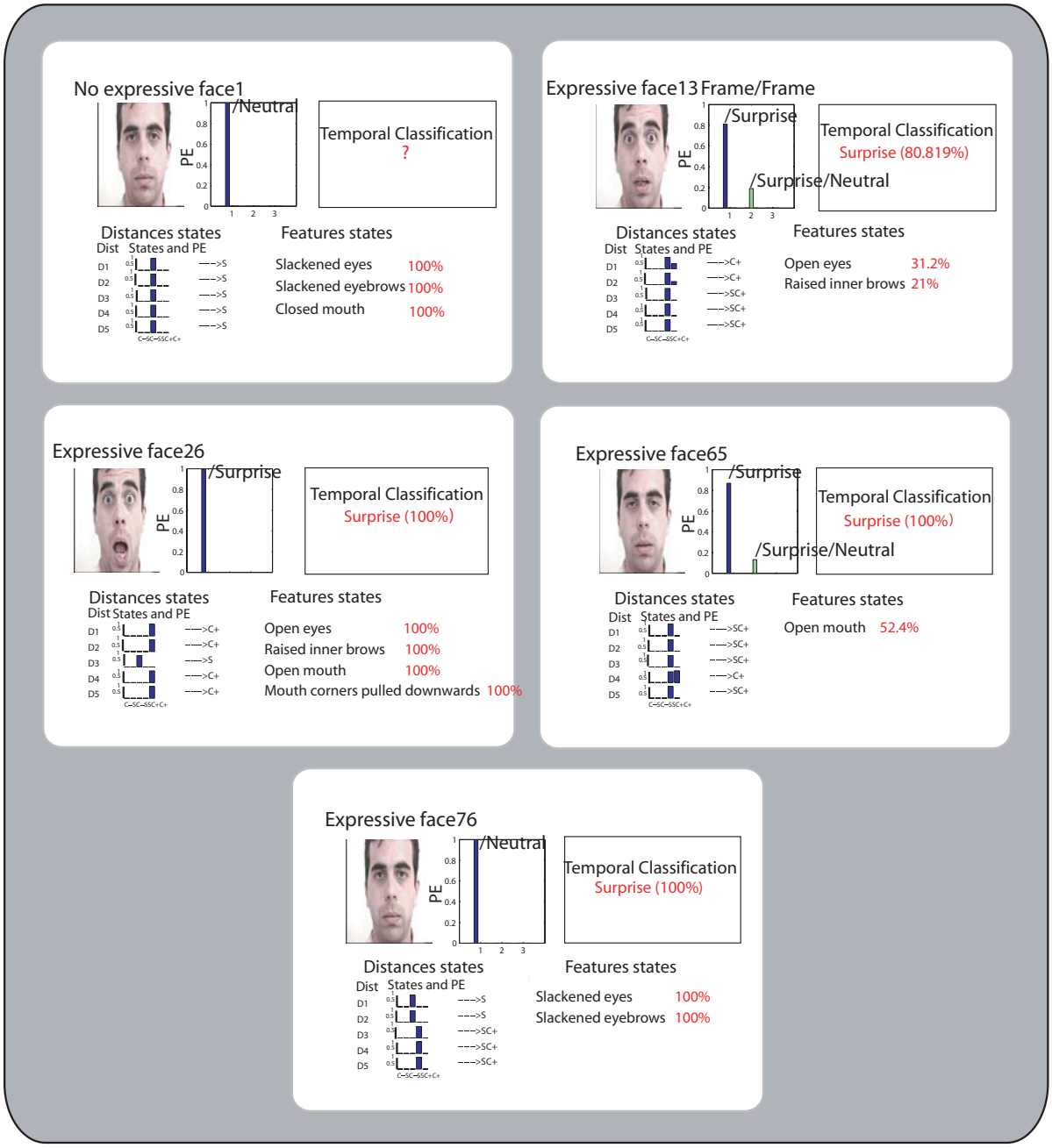

Figure 4.11: Examples of dynamic classification results during a *Surprise* expression.

4.5 Conclusion

In this chapter we have presented our preliminary works on the introduction of the temporal information for the analysis and classification of facial expressions.

Firstly, the temporal information has been added to enhance the static classification based on automatic segmentation. This Frame-by-Frame classification leads to the improvement of the performances compared with those obtained on the static classification (in order to demonstrate the efficiency of our Frame-by-Frame classification of facial expression, we compare it with a HMM based classifier (see Appendix 7.2))

Secondly we propose a method which takes into account all the available information during the expression sequence for the classification. The accuracy and robustness of our approach lies in the modeling of the uncertainties and dynamics of facial features behavior. The efficiency of our approach is achieved through the combination of the whole set of facial features deformations from the *beginning* until the *end* of the expression sequence.

The two methods are based on the TBM and then share the same properties as the static classification: its ability to deal with imprecise data, to handle with unknown expressions and finally to model the doubt between some expressions. However similarly to our work on static classification which has been evaluated on a large set of data from different databases, the results of the Frame-by-Frame and Dynamic classification need to be evaluated on more data to go further in their evaluation.

There are three limitations in our approach to automatic facial expression analysis. First, applied on data obtained from our automatic segmentation, our method is, in some extent, sensitive to head motion. This may limits the application of our system to non-mobile people. Second, since feature displacements are measured with respect to their *Neutral* positions, the knowledge about someone's facial features in the *Neutral* facial expression has to be acquired prior to analyzing facial expressions. Third, the behavior of the system has to be evaluated in the case where the person speaks at the same time in order to know how the lips motion affects the facial features analysis.

Compared with the existing works on facial expressions analysis and in addition to the properties of our system pointed out in the static classification (see section 3.9), our approach enjoys several favorable properties: at each frame, the system gives an analysis of the facial features states and their associated piece of evidence even in the case of *Unknown* expression; it manages to deal with asynchronous facial features deformations; it is able to recognize entire facial expressions sequences.

Part III

Towards bimodal (audio-video) classification system

Préambule. Dans le but d'étendre le système proposé pour la classification des expressions vers un système de classification bimodale, nous avons initié une étude sur la modalité vocale. Nous concluons cette étude par une discussion sur le problème de la fusion des deux modalités (faciales et vocales).

Preamble. In order to extend the proposed system for expressions classification towards a bimodal classification system, we have initiated a study of the vocal modality. Based on our works on facial and vocal expressions, we also discuss the fusion of these two modalities.

Chapter 5

Vocal expressions classification

This work has been realized in the scope of a PhD exchange program in the Network of Excellence SIMILAR [Sim04] in collaboration with the laboratory of *Théorie des Circuits et Traitement du Signal* (TCTS) in Mons in Belgium, in the research group of Professor Thierry Dutoit, in collaboration with Baris Bozkurt, Laurent Couvreur and Devrim Unay. In the following I present the result of our collaboration [Ham05a], [Ham05b].

5.1 Introduction

From the psychological point of view, Charles Darwin [Dar72] is the first author who was interested in the systematic study of expressions corresponding to emotions. Its observations took into account facial expressions as well as vocal expressions, produced by human beings belonging to various cultures. During the 20th century numerous studies focused on facial expressions. Comparatively, the study of vocal expressions was longer neglected. However, the user interface for computer systems is evolving into an intelligent multimodal interface by taking into account several modalities such as the user's behavior, speech and facial expressions in order to make the use of machines as much natural as possible. Consequently, several works were carried out on vocal expressions recognition. The vocal expressions analysis includes two sub-problem areas: specification and analysis of the characteristic features to be estimated from the vocal signal and classification of the extracted features into one of the predefined vocal expression.

5.1.1 Vocal characteristics analysis

It consists in the study of the encoding process of the emotion in the voice. It describes the effect of the emotional states on a whole set of vocal characteristics. The most usual measurements are the parameters derived from the fundamental frequency, and the acoustic intensity.

- *The fundamental frequency* (Pitch or F0), expressed in hertz (Hz), corresponds to the number of repetitions per second of the fundamental period of the acoustic signal;
- *The acoustic intensity*, expressed in decibels (dB), is derived from the amplitude of the acoustic signal.

Based on the contours of F0 and the intensity some researcher have analysed the evolution of these two parameters during an expression. A set of statistical measurements are then defined for the analysis process. In addition to F0 and the energy, some aspects of the duration of the expressions were also reported in various studies. The most current measurements are the speech rate (for example the number of syllables pronounced per minute), the proportional duration of the pauses according to the total duration of the expressions and the number of pauses. We can also found the analysis of the formants noted F1, F2, F3, and F4 which are defined as areas of the acoustic spectrum in which energy is particularly high, reflecting the resonances produced by the shape of the vocal tract. Some authors used also the duration of the voiced parts according to the not-voiced parts of the speech signal. Banse & Scherer [Ban96] included a whole set of measurements derived from the average spectrum in the long run (long term average spectrum). Juslin & Laukka [Jus01] included the short term disturbances of F0 ("Jitter") which corresponds to fast and random fluctuations of the duration of open/closed of the vocal cords. Moreover they evaluate the average height of the formants and the frequency bandwidth which contains energy associated with the formants.

Based on all these results, the main and most recent vocal expressions analysis has been performed by Juslin [Jus03] and Scherer [Sch03a] to identify the behaviors of all these parameters during a vocal expression. The analysis of Scherer [Sch03a] focuses on 6 expressions (*Stress*, *Anger*, *Fear*, *Sadness*, *Joy* and *Boredom*). Their analysis is based only on the 7 frequently used acoustic parameters described in Table5.1. The results of their study show a distinction between two sets of expressions: *Anger*, *Fear* and *Joy*, on the one hand, and the expressions of *Sadness* and *Boredom*, on the other hand.

	Vocal expressions					
	Stress	*Anger*	*Fear*	*Sadness*	*Joy*	*Boredom*
Intensity	↗	↗	↗	↘	↗	.
F0 floor/mean	↗	↗	↗	↘	↗	.
F0 variability	.	↗	.	↘	↗	↘
F0 range	.	↗	↗ (↘)	↘	↗	↘
Sentence contours	.	↘	.	↘	.	.
High frequency energy	.	↗	↗	↘	(↗)	.
Speech and articulation rate	.	↗	↗	↘	(↗)	↗

Table 5.1: Synthesis of the results of Sherer study [Sch03a]. ↗: increase, ↘: decrease.

In their analysis Juslin and Laukka [Jus03] studied *Anger*, *Fear*, *Sadness*, *Joy* and *Tenderness* expressions. In addition to the acoustic parameters of Scherer they also studied 7 more rarely used parameters (régularités microstruct, proportion de pauses, précision articulation, Formant 1 (height and width), Jutter, glotal wave form). The authors found the same expressions confusion as Scherer but the less frequently used parameters exhibit different patterns for the 5 considered emotional categories. However, the results for these parameters are very few and are not always identical for all the subjects and then more results are required to confirm them.

Based on these results we can assume that if we assume that the usually studied expressions of *Joy*, *Anger* and *Fear* correspond to strongly activated emotion, the usually measured acoustic parameters reflect the *emotional activation*. The emotions which include a strong

activation (*Joy*, *Anger*, *Fear*) are positively correlated with the statistical parameters of F0, the intensity and the speech rate. Whereas the states which include a small degree of activation (*Sadness* and *Tenderness*) are negatively correlated with the values of F0, intensity, as well as the speech rate. These conclusions are also recently studied and confirmed by the work of Schröder [Sch03b].

5.1.2 Vocal expressions: towards bimodal classification

In this section we present studies dealing with the processes of vocal expressions decoding. The main goal of these works is the discrimination of several predefined vocal expressions (from 4 to 10 different classes) using a set of predefined acoustical parameters (see section 5.1.1). In the following we will survey current state of the art in the vocal expressions recognition and its use towards bimodal (voice + video) expressions classification. However, we only focus on the recently proposed methods which had the greatest impact.

5.1.2.1 Vocal expressions classification

Petrushin [Pet00] proposed an automatic system for vocal expressions recognition. Their study deal with 700 short utterances (sentences) told by 30 subjects expressing five emotions: *Happiness*, *Anger*, *Sadness*, *Fear* and *Normal* (unemotional). The classification is based on some statistics (mean, standard deviation, minimum, maximum and range) for a set of acoustical variables: the pitch (F0), the first three formants (F0, F1 and F3) and their bandwidths, the energy and the speach rate. 14 of these features are selected using the RELIEF-F algorithm [Kon94]. The classification is made using a set of neural networks trained on different subsets of the training set using bootstrap aggregation and cross-validated techniques. The obtained classification rates are: $55-75\%$ for *Happiness*, $60-70\%$ for *Anger*, $70-80\%$ for *Sadness* and $35-55\%$ for *Fear*.

In order to class the six universal emotional expressions plus *Neutral* expression, Nogueiras [Nog01] proposed an HMM classifier. The study is based on the INTERFACE Emotional Speech Synthesis database [INT00], recorded in four different languages from two actors (one male and one female). In order to better discriminate between the studied expressions the authors try a novel set of statistical parameters on the pitch (maximum of auto-correlation and its first and second order derivatives and the first and second order derivatives of the logarithm of the pitch) and for the energy (first and second derivatives of the logarithm of its low pass filtered value). Based on these parameters seven HMM are trained on a part of the recording and tested on the remaining part. The obtained classification rates on 555 utterances test set are: 73% for *Happiness*, 80% for *Anger*, 80% for *Sadness* and 81% for *Fear*, 90% for *Surprise*, 80% for *Disgust* and 77% for *Neutral*. A confusion appears between two sets of expressions: (*Surprise*, *Hapiness*, *Anger*) and (*Fear*, *Disgust*, *Sadness*).

Recently Ververidis [Ver04] proposed to enhance the classification rate by taking into account the gender information. A total of 87 features has been computed over 500 utterances of the Danish Emotional Speech (DES) database [Eng96]. The sequential forward selection method based on the cross validated correct classification rate of Bayes classifier has been used in order to discover the $5-10$ statistical features (on the pitch, the energy, the formants F1, F2, F3 and F4) able to classify the samples in the best way for each gender. The obtained classification rates are: 54% for *Happiness*, 57% for *Anger*, 58% for *Sadness*, 61% for *Surprise* and 55% for *Neutral* on the female subjects and 43% for *Happiness*, 56% for *Anger*, 80% for

Sadness, 60% for *Surprise* and 67 for *Neutral* on the male subjects. A confusion also appears between (*Surprise*, *Happiness*, *Anger*) expressions and (*Fear*, *Disgust*, *Sadness*) expressions.

Some other works have focused on the discrimination between vocal expressions [Sat01], [Lee04]. The main difference between them is the number of the acoustical characteristics and the statistical measures used to separate between the considered expressions.

5.1.2.2 Bimodal expressions classification

Several researchers have been interested in combining the voice modality with the face modality for a better discrimination of the emotional state.

Visual information modifies the perception of speech [Mas98]. In order to know if the same conclusion can be done on the emotion recognition, De Silva *et al* [De 97] conducted an experiment. 18 people were required to recognize emotion using visual and acoustic information separately from audio-visual database recorded from two subjects. They concluded that some emotions are better identified with audio such as *Sadness* and *Fear*, and others with video, such as *Anger* and *Happiness*. These conclusions have been also confirmed by Chen *et al* [Che98] which showed that audio and visual modalities give complementary information, by arguing that the performances of the system increased when both modalities were considered together. Despite these conclusions and compared to the recent advances in unimodal facial and vocal expressions analysis, there are few studies on audio-visual expressions classification systems. A detailed review of the recently proposed methods can be found in [Pan03] and [Seb04].

Chen *et al* [Che98] bimodal classifier was based on rule-based method for the classification of the six universal emotional expressions. From the speech signals a set of acoustic features, pitch, intensity, and pitch contours were estimated for the vocal classification process. From the visual signal, lowering and rising of the eyebrows, opening of the eyes, stretching of the mouth, and presence of a frown, furrow and wrinkles were manually measured from the input images for the visual classification process. The used database corresponds to 36 video clips of Sinhala speakers which portray each of the studied expressions six times using both vocal and facial expressions. Based on the features (vocal and facial) extracted from this data a set of rules were defined for the classification process. However, the evaluation of the defined rules has been done only on the mentioned data set. It is not known whether or not the defined rules are suitable for emotion recognition from audiovisual data of an unknown subject.

De Silva and Ng [De 00] proposed also a rule-based method based on audio-visual emotion recognition in which the outputs of the unimodal classifiers are fused at the decision level. From the speech signal only the pitch and its contours were estimated [Med91]. From the face signal, the optical flow method proposed in [Bla93] was used to detect the motion of the mouth corners, the top and the bottom of the inner corners of the eyebrows. The used database corresponds to 144 2-s-long video clips of two English speakers which has been asked to portray 12 intensities for each emotion displaying the related prototypic facial expression while speaking a single English word of his choice. The pertinent audio and visual material has been processed separately. The classification was done by a nearest-neighbor method for the facial expressions and HMM-based method for the vocal expressions. Per subject, the results of the two classifications have been plotted in a graph. Based on the two resulting graphs, the authors defined the rules (by expertise) for emotion classification of the audiovisual data. The obtained recognition rate is equal to 72% for a reduced input data set (ie. 10% of the input samples for which the used rules could not yield a classification into one of the emotion

categories were excluded from the data set). It is not known, therefore, the precision of the proposed method for unknown subjects.

Yoshitomi *et al* [Yos00] also proposed a method for the classification of audio-visual data into five expressions: *Happiness*, *Sadness*, *Anger*, *Surprise*, and *Neutral*. The used acoustic features are the pitch, the intensity, and the pitch contours. Based on these features HMM-based method is used for the vocal expressions classification. Facial features are captured using a visible rays (VR) camera and an infrared (IR) camera at the same time. From each one, the two frames corresponding to the maximum intensity of the speech are selected to detect mouth, eyes and eyebrows bounding boxes. Compared to the corresponding regions in the *Neutral* state a differential image is defined. Then a discrete cosine transform has been applied to yield a VR and an IR feature vector used by two separate neural networks for facial expressions classification. Their results were further summed to those obtained with vocal modality to decide the final output category. The analysis was based on 100 video clips of one female Japanese professional announcer. She was asked to pronounce a Japanese name "Taro" while portraying each of the five emotions 20 times. The reported recognition rate is equal to 85% for a reduced input data set (i.e., 34% of the input samples for which the proposed method could not yield a classification into one of the emotion categories were excluded from the data set). Similarly to Chen *et al* and De Silva and Ng, it is not known whether and with which precision this method could be used for emotion classification of audiovisual data from an unknown subject.

Chen and Huang [Che00] proposed a classification of the six universal emotional expressions based on the analysis of 180 video sequences of five subjects. Each subject displayed each expression six times by producing the appropriate facial expression before or after pronouncing a sentence with the appropriate vocal emotion [Coh03b]. Facial features motion tracking (eyes, eyebrows and mouth) [Tao99] is done to obtain a vector containing the strengths of the facial muscle actions. A naïve Bayesian classifier is then used for the classification process. From the speech signals, pitch, intensity and speech rate have been used for the vocal expressions classification using a Gaussian distribution of each expression. Given that in each used video clips a pure facial expression occurs before or after a sentence spoken with the appropriate vocal emotion, the authors applied the two methods (two modalities) in a sequential manner according to whether the subject is speaking or not. In person-dependent experiments, half of the available data have been used as the training data and the other half as the test data. A 79% average recognition rate has been achieved in this experiment. In person independent experiments, data from four subjects have been used as the training data, and the data from the remaining subjects have been used as the test data. A 53% average recognition rate has been reported for this experiment.

Zeng *et al* [Zen05] proposed a method for the classification of the 6 universal emotional expressions plus the *Neutral* expression plus 4 cognitive states (*Puzzlement, Interest, Boredom* and *Frustration*). From the visual signal a tracking process [Tao99] is used to detect 12 predefined facial features motion according to the *Neutral* state. Pitch and the energy acoustic features are detected from the speech signal and normalized according to the *Neutral* state values. Each one of the facial and the vocal features were quantized into 19-size codebook by vector quantization. One HMM is defined for the classification of each feature (facial features, pitch and energy). For integrating coupled audio and visual features they proposed multi-stream fused HMM for the classification process based on the combination of the 3 HMM. The used database include 20 subjects (10 males and 10 females). Experimental results based on the analysis of 11 affect states of 20 subjects (repeated 20 times, at each time all the

sequences of one subject are used as a test sequence and the remaining 19 subjects for the training) lead to a recognition rates of 38.64% based only on the face, 57.27% based only on the pitch, 66.36% based only on the energy and 80.61% based on their combination.

The methods described above fused the two modalities either at the decision-level, or at the feature-level. In their works Busso *et al* [Bus04] try to identify the advantages and limitations of unimodal systems, and to show which fusion approaches are more suitable for emotion recognition. Four emotional expressions are analysed: *Sadness*, *Happiness*, *Anger* and *Neutral*, using a database recorded from an actress who reads 258 expressive sentences. The visual information corresponds to the expressive facial motion captured with markers attached to the face. The vocal information corresponds to a set of pitch statistics [Boe01] and to the voiced/unvoiced speech ratio. To fuse the facial expression and acoustic information, two different approaches were implemented and compared: feature-level fusion and decision level fusion using a fusing criteria. The overall performances of both approaches were similar. However, the recognition rate for specific emotions presented significant discrepancies. In the feature-level bimodal classifier, *Anger* and *Neutral* state were accurately recognized compared to the facial expression classifier, which was the best unimodal system. In the decision-level bimodal classifier, *Happiness* and *Sadness* were classified with high accuracy. The authors conclude that the best fusion technique will depend on the application.

Reference	Classification	Nbr of expressions	Nbr test subjects	Performances (%)
Vocal expressions classification approaches				
[Pet00]	neural network NN	5	30	70
[Nog01]	HMMs	7	2	80
Ververidis [Ver04]	Bayesian classifier	5	4	Male: 61.1 Female: 57.1
Bimodal (vocal + facial) expressions classification approaches				
[Che98]	rules based method	6	-	-
[Yos00]	neural network NN	5	1	85
[De 00]	rules based method	6	2	72
[Che00]		6	5	53 - 79
[Bus04]		4	1	89
[Zen05]	multi stream fused HMM	7	20	80.61

Table 5.2: Comparisons of vocal expressions and bimodal expressions recognition systems.

5.1.3 Conclusion

The different systems are very difficult to compare because they are evaluated on different databases and they do not study the same vocal expressions. However, similarly to the visual analysis the same seven expressions are commonly considered to characterize human emotional states in the vocal signal (*Happiness*, *Surprise*, *Anger*, *Sadness*, *Disgust*, *Fear* and *Neutral*). However, contrary to the Ekman's work on facial expressions, universality of the vocal expressions has not been proved. Some efforts have been made to discriminate between them. Based on the most used prosodic features of the speech signals (pitch, energy, and speech rate) a lot of statistical measures have been defined from these features to characterize and separate between the studied expressions. Even with a large set of statistical features the proposed approaches show a great confusion between *Anger*, *Surprise* and *Happiness* on

the one hand and *Neutral* and *Sadness* on the other hand. This confusion is also found when humans try to make the same classification on the used databases.

It has also been shown that the performances of emotion recognition systems can be improved by the use of multimodal information. However, it is not clear which technique is the most suitable to fuse these modalities, if these information lead to a separation between the confused classes and finally if it is justified to look for the same expressions in the voice as in the face information.

The final aim of our work is a bimodal expressions classification system based on the face and voice modalities. For the face modality we will use our works on facial expressions. In the following we consider the vocal expressions analysis and classification. However, contrary to the common approaches which try to solve the confusions between two sets of vocal expressions by the addition of new characteristics, we rather consider this conflict not as a criterion of dissimilarity but as a criterion of similarity between the confused expressions. Finally we will present our perspectives to fuse the two modalities for a bimodal expressions classification system.

Section 5.2 presents the speech database used in this work. In section 5.3 we present features extraction and analysis. Section 5.4 is dedicated to the results and discussion.

5.2 Speech database

For our experiments, we used the DES database [Eng96]. The data were collected from two male and two female professional actors. The following expressions have been investigated : *Neutral*, *Surprise*, *Happiness*, *Sadness* and *Anger*. For each expression, there are 2 single words, 9 sentences and 2 longer passages of continuous speech. A high quality microphone was used, which did not influence the spectral amplitude or phase characteristics of the speech signal. To check the accuracy of the simulated data, a listening test has been performed by the authors of the database to check if listeners (20 normal-hearing, 10 of each gender) could identify the emotional content of the recorded utterances. The utterances were correctly identified with an average rate of 67% (see Table 5.3). *Surprise* and *Happiness* were often confused as well as *Neutral* and *Sadness*.

Expert \ Human	*Neutral*	*Surprise*	*Happiness*	*Sadness*	*Anger*
Neutral	**60.8**	2.6	0.1	31.7	4.8
Surprise	10.0	**59.1**	28.7	1.0	1.3
Happiness	8.3	29.8	**56.4**	1.7	3.8
Sadness	12.6	1.8	0.1	**85.2**	0.3
Anger	10.2	8.5	4.5	1.7	**75.1**

Table 5.3: Confusion matrix from subjective human evaluation [Eng96]. Columns represent the vocal expression selected for utterances for the vocal expressions input of each row.

5.3 Features extraction and analysis

Based on recent studies of vocal expressions analysis, several prosodic features have been defined. Guided by the works of [Sch03a], [Jus03], [Pet00], [Ver04] we restrict ourselves to the following features: the pitch, the energy, the SPI (Soft Phonation Index) and the speech rate. Analysis is carried out to extract the pitch, the energy and the SPI extracted from speech frames being part of recording with constant length (30msec) and constant shift (10msec). Next we perform a statistical analysis in order to select the acoustical parameters that could display the differences between vocal expression categories. Our analysis was guided by the works [Sch03a], [Jus03], [Pet00] to realize the correlation between the sets of characteristics extracted for each parameter and the vocal expressions. In the following we are interested only in the normalized characteristics (zero-mean and standard deviation to 1) presented in Table 5.4.

	Speech rate					
	Rang	Median	Standard deviation	Rises	Falls	Max
F_0	X	X	X	X	X	X
Energy	X	X	X	X	X	X
SPI	–	–	–	–	–	X

Table 5.4: Statistical parameters used for each characteristic.'X': used, '–': not used.

5.3.1 Pitch

The pitch (F0) is the fundamental frequency of the acoustic signal. This feature is computed using an autocorrelation based pitch estimator [Qua01]. Statistics related to F0 such as minimum, maximum, mean, median, range, standard deviation are computed. Flatness of intonation is also measured with two values: median values of the rises and falls of F0 [Sch03b]. Figure 5.1 presents the result of the range (maximum-minimum), median and standard deviation for F0. The value of each bar corresponds to the mean value for all the data for each expression. The standard deviation of this value is also reported. Figure 5.2 presents the median of the rises and of the falls of F0 for every expression. Figure 5.1 and Figure 5.2 show that two groups of expressions appear: the statistical values of F0 for *Surprise*, *Anger* and *Happiness* examples are comparatively higher than the corresponding values for *Neutral* and *Sadness* examples.

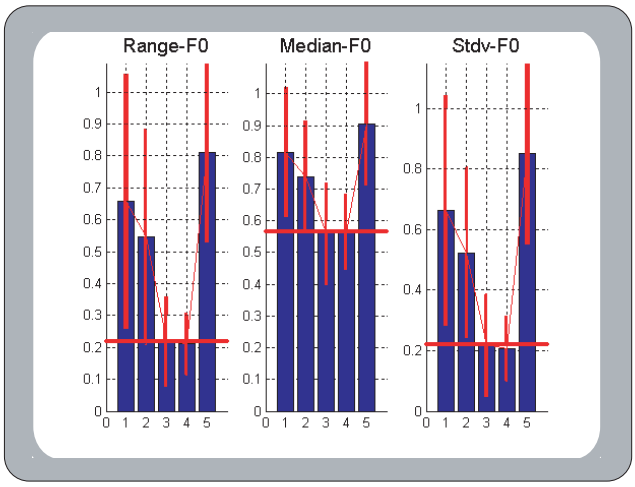

Figure 5.1: Mean values of range, median and standard deviation of F0 for all the data and all the expressions. The bars represent the expressions in the following order : 1) *Anger*, 2) *Happiness*, 3) *Neutral*, 4) *Sadness*, 5) *Surprise*.

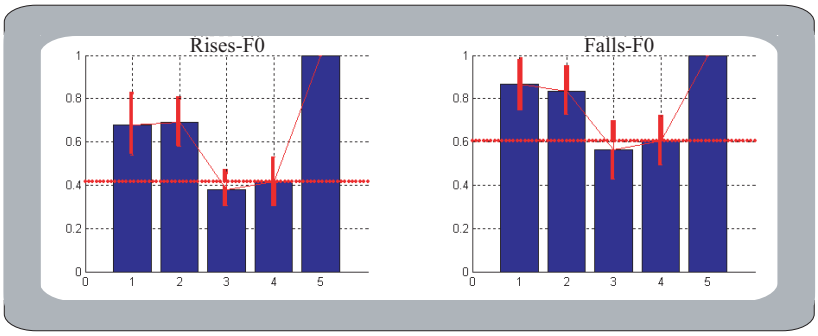

Figure 5.2: Mean values of rises and falls for F0 for all the data and all the expressions. The bars represent the expressions in the following order : 1) *Anger*, 2) *Happiness*, 3) *Neutral*, 4) *Sadness*, 5) *Surprise*

5.3.2 Energy

The signal energy is computed (in decibels) as the sum of square of the discrete signal [Qua01]. We compute the energy of only the voiced segment in utterances to avoid jumps at plosives. Similarly as applied to the pitch, we compute a set of global statistics such as minimum, maximum, median, range, standard deviation and medians of rises and falls. Figure 5.3 and Figure 5.4 present the statistical characteristics of energy. *Sadness* and *Neutral* speech show lower range, median, standard deviation, rises and falls compared to the other vocal expressions. These results are coherent with the fact that *Anger*, *Happiness* and *Surprise* require more energy than *Neutral* and *Sadness* expressions [Sch03b].

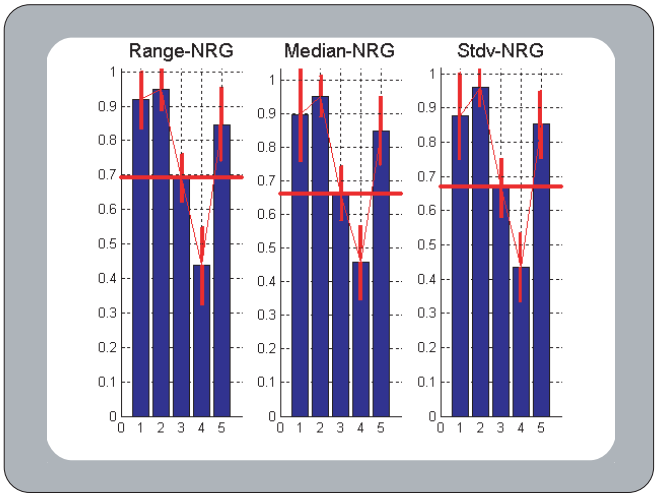

Figure 5.3: Mean values of range, median and standard deviation of energy for all the data and all the expressions. The bars represent the expressions in the following order : 1) *Anger*, 2) *Happiness*, 3) *Neutral*, 4) *Sadness*, 5) *Surprise*.

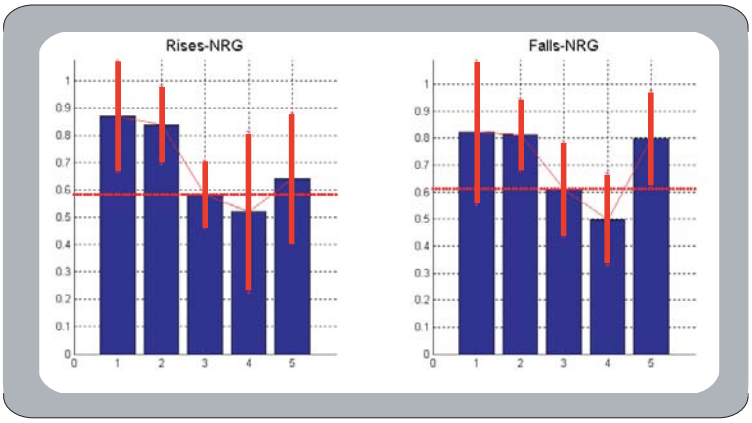

Figure 5.4: Mean values of rises and falls for energy for all the data and all the expressions. The bars represent the expressions in the following order : 1) *Anger*, 2) *Happiness*, 3) *Neutral*, 4) *Sadness*, 5) *Surprise*.

5.3.3 SPI

SPI is a spectral measure of the ratio of low-frequencies (70-1600HZ) to high-frequencies (1600-4500Hz) [Del93]. It is used as a simple approximation of the *harshness* vs. *softness* of the voice quality in the area of speech therapy. The analysis of the characteristics of the SPI of voiced frames shows that only the maximum value is relevant for classification. This value presents the same behavior as F0 and the energy (Figure 5.5).

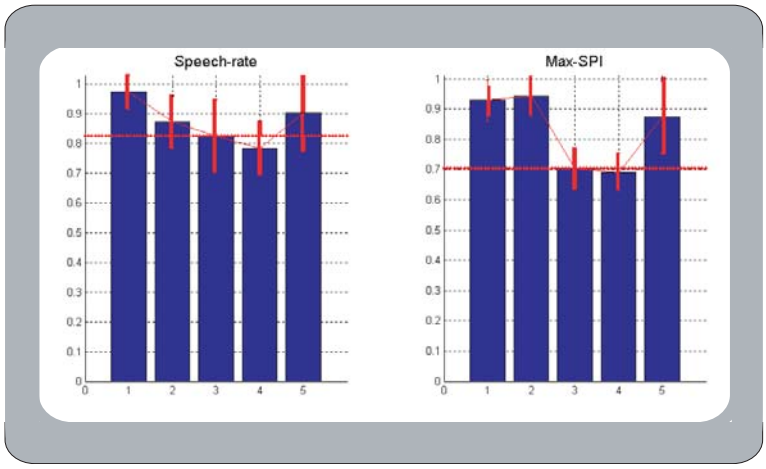

Figure 5.5: Speech rate mean values and SPI maximum mean values for all the data and all the expressions. The bars represent the expressions in the following order : 1) *Anger*, 2) *Happiness*, 3) *Neutral*, 4) *Sadness*, 5) *Surprise*.

5.3.4 Speech rate

The Speech rate is computed for each recording as the number of phonemes spoken in a given time interval. Since the text content is known, the number of phonemes of each recording is available in the database. Only the estimation of speech duration from the recorded signal is necessary for this task. Recordings are segmented into speech and non-speech segments by applying an energy threshold for decision. The energy threshold is defined proportionally to the mean energy of each recording. The analysis of the speech rate (Figure 5.5) shows that this feature is higher for *Surprise*, *Anger* and *Happiness* than for *Neutral* and *Sadness*.

5.3.5 Conclusion of the analysis

Expressions such as *Anger*, *Surprise*, *Happiness* have higher values of F0, energy, SPI and speech rate which means that they are related to a strong activity. On the contrary, expressions such as *Sadness* or *Neutral* present smaller values of F0, energy and SPI as well as a decrease of speech rate, which means that they are related to a small activity. These observations lead us to conclude that the frequently used acoustic parameters do not discriminate the 5 considered vocal expressions. These parameters exhibit two groups of expressions: *Anger*, *Surprise* and *Happiness*, on the one hand, and *Sadness* and *Neutral*, on the other hand.

5.4 Results

5.4.1 Five expressions classification

The analysis of section 5.1.1 shows that there are prosodic similarities between several vocal expressions. We first want to confirm these results and to see the effect of these characteristics on the discrimination between the vocal expressions. We use the 12 global acoustic features described in section 5.1.1 (median, range, standard deviation, rises and falls for F0 and energy, maximum for SPI and speech rate) with a Bayesian classifier to classify the five vocal expressions. In order to minimize the effect of the lack of data we use a bootstrap method [Kal02] to better estimate the classification rates. It consists in duplicating the number of data by random pulling with handing-over. In our case 50 databases are built from the initial complete database. Classification rates are calculated in the following way: at each iteration, we train the classifier on one database and we test on the initial complete database. The process is reiterated on the 50 databases. The final rates are the mean of the 50 rates. Table 5.5 presents the results of the classification. To test the validity of the chosen characteristics for the vocal classification process on the same data we compare these results with those of Ververdis *et al* [Ver04]. The classification is based on the same database DES [Eng96]. Ververidis *et al* used others additional characteristics (the pitch, the energy, the formant F1, F2, F3 and F4) and a set of $5-10$ statistical features for each one. The classification rate obtained in [Ver04] is around 50% (*Neutral* (51%), *Surprise* (64%), *Happiness* (36%), *Sadness* (70%) and *Anger* (31%)) whereas ours is around 54%. The average of the two classification rates are comparable while the classification rates of the expressions separately is very different. A more important observation is that our results are more homogeneous: the classification rate is almost the same for all the expressions which is not the case in [Ver04]. Moreover our rates are closer to those of the human classifier (see Table 5.3).

Expert \ System	*Neutral*	*Surprise*	*Happiness*	*Sadness*	*Anger*
Neutral	**46.76**	23.92	12.26	3.3	13.73
Surprise	20.11	**51.69**	6.5	5	16.61
Happiness	7.11	5	**56.61**	24.69	6.5
Sadness	4.57	3.19	28.76	**61.80**	1.65
Anger	12.5	29.11	4.26	1.84	**52.26**

Table 5.5: Confusion matrix with a Bayes classifier.

5.4.2 Passive versus Active classification

Comparison with Table 5.3 indicates that human listeners show the same tendency. The emotions that have been confused are those with similar acoustic characteristics (see section 5.1.1). Considering the confusions as indicators of the similarity perceived between the confused expressions we decide to create two classes: *Active* which includes *Anger*, *Happiness* and *Surprise* and *Passive* which includes *Neutral* and *Sadness*. To be sure of the discrimination between these two new classes, we compare the classification rates obtained with 4 different classifiers: the Bayesian classifier, the Linear Discriminant Analysis (LDA) [Dud01], the K

nearest neighbours (KNN) with 5 neighbours and Euclidian metric [Dud01] and the Support Vector Machine with gaussian radial basis function kernel (SVM) [Bur98]. The classification rates are obtained by 5-fold cross validation. The results of classification (Tables 5.6-5.7) show that the recognition rates of Bayesian classifier and LDA are lower than SVM and KNN. This is due to the fact that Bayesian classifier assumes Bayesian distributions of classes, which may be a false assumption for our dataset, and LDA performs a linear separation while our data may be non-linear. The KNN performs better result than LDA, however SVM gives the best classification rates (Tables 5.6 right - 5.7 left). The presented results (Table 5.7) makes it possible to confirm that the chosen characteristics are sufficient for our two classes classification.

	Bayesian classification		LDA classification	
System / Expert	*Active*	*Passive*	*Active*	*Passive*
Active	**78.84**	21.15	**96.79**	3.2
Passive	19.23	**80.76**	46.15	**53.85**

Table 5.6: Left: results of Bayesian classification; right: results of LDA classification.

	KNN classification		SVM classification	
System / Expert	*Active*	*Passive*	*Active*	*Passive*
Active	**83.33**	16.67	**89.74**	10.26
Passive	11.54	**88.46**	13.46	**86.54**

Table 5.7: Left: results of KNN classification; right: results of SVM classification.

5.5 Conclusion

In order to integrate speech modality to expressions classification system based on video, we investigated acoustic properties of speech associated with five different vocal expressions. The analysis of the acoustic features enables to note that the considered acoustic features provide rather limited support to separate the five vocal expressions. However results show that grouping expressions into two larger classes *Active* versus *Passive* according to the statistical parameters derived from acoustic features results in successful classification. The same confusions are found for a classification by humans, which leads us to define two classes of vocal expressions: *Active* and *Passive*. The interest of this classification is that it is more compliant with real applications (for example as call center). The development of a multimodal expressions recognition system is under study. Such a system will combine at the same time both modalities (video and speech) for better recognition or will use them separately according to the context of the application.

Facial expressions and vocal expressions systems have been done largely independent of each other. The mentioned works in facial expression recognition used still photographs or video sequences where the subjects exhibit only facial expression without speaking any words.

Similarly, in the proposed vocal expressions classification systems the detection is only based on the audio information. However, in some situations people would speak and exhibit facial expressions at the same time. Thus bimodal emotional expressions classification system has to be developed to handle with such cases.

5.6 Discussion: towards bimodal emotional expressions classification

Human makes use of more than one modality to recognize emotions [Meh68]. Then it is expected that similarly, the performances of multimodal systems will be higher than automatic unimodal systems.

In our work, the first step to be fulfilled is the combination of the two modalities (face and speech) towards a bimodal system for expressions recognition. However, one major problem of this combination process is that we have not the same expressions classes for the two modalities. Moreover, the combination needs the definition of an integration model of the two modalities: the fusion of the classification results of the different modalities (Figure 5.6 dashed line) or the fusion of all the features before the classification process (Figure 5.6 plain line).

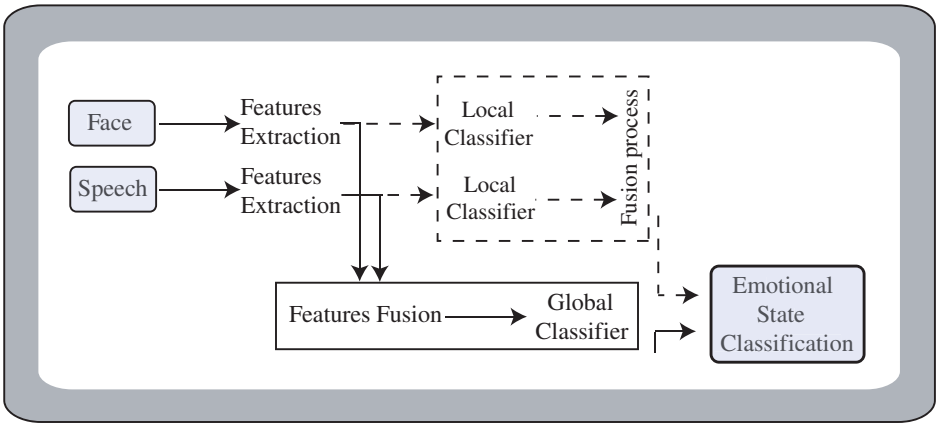

Figure 5.6: Bimodal recognition scheme of human emotional state.

However according to Pantic [Pan03] the first solution which is the most commonly used is almost certainly incorrect. People display audio and visual communicative signals in a complementary and redundant manner. De Silva *et al* [De 00] and Chen *et al* [Che98] efforts proved that the two modalities are complementary for a human emotional state detection. Then for a human like bimodal analysis face and voice signals cannot be considered mutually independent and cannot be combined in a context-free manner at the end of the intended analysis but, on the contrary, the input data should be processed in a joint feature space and according to a context-dependent model. Busso *et al* [Bus04] claimed that the most suitable fusion technique will depend on the application.

Moreover, based on automatic voice and face detection features, noisy and partial input data can appear. A bimodal fusion system should be able to deal with these imperfect data and generate its decision so that the certainty associated to it varies in accordance

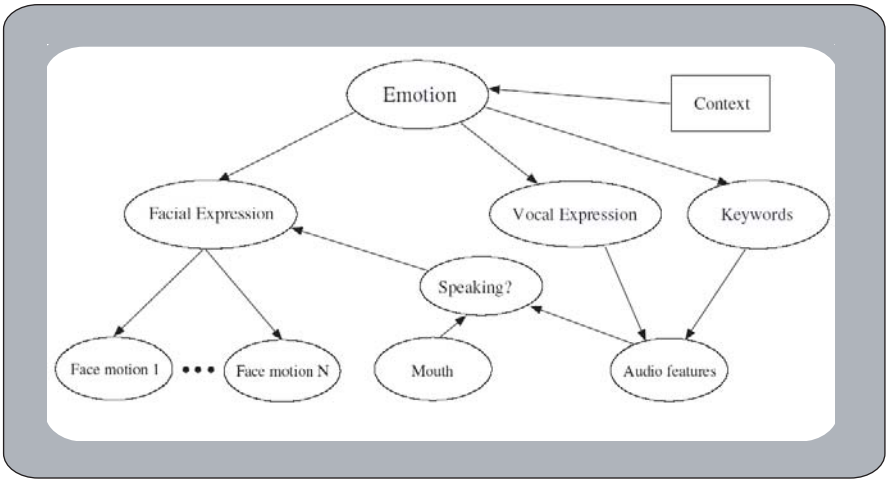

Figure 5.7: Bayesian network topology for bimodal emotion expression recognition [Seb04].

to the input data. According to Pantic *et al* [Pan03] a way to achieve this is to consider the time instance (the current observed data) versus the time scale (previously observed data) dimension associated to human nonverbal communicative signals. Sebe *et al* [Seb04] emphasize the suitability of probabilistic graphical models such as HMM, Bayesian networks and Dynamic Bayesian networks for fusing such sources of information. They also proposed a Bayesian network topology (see Figure 5.7) to combine the two modalities in a probabilistic manner. In their model they also added the context of the application (if this one is available) and the speech variable that indicates whether a person is speaking or not.

Based on our previous work on facial and vocal expressions classification, we can propose the same model as Sebe *et al*'s one but using the TBM rather than the Bayesian model to combine our dynamic facial expressions classification results and our vocal expressions classification results. This approach proves to be more adapted to the problem of facial expressions classification and proves to be able to deal with imprecise data and to model a doubt between some expressions. Indeed, most of the time, people show blends of emotional expression. Thus automatic bimodal expressions classification has to deal with mixture between expressions.

There are two additional problems: first the different features formats and timing and secondly the different classes associated to the two modalities. A potential way to achieve such a bimodal system is to develop context-dependent versions of the TBM. To do this we can potentially learn application-dependent, user-dependent, and context-dependent rules by watching the user's behavior in the sensed context [Pen00].

In everyday life people communicate through multiple modalities: their face, their speech and their body. Hence a system that attempts to interact with humans should be able to extract and interpret all these cues so as take into account their dynamic behavior and their emotional state. An ideal analyzer of human nonverbal affective feedback (see Figure 5.8) should generate a reliable result based on multiple input signals acquired by different sensors. In addition to facial features and voice information, other modalities have to be added such as gaze estimation (see Appendix A) and head and body gesture estimation (posture). Then this system can be successfully integrated to different applications: interactive games with the ability to adapt to the user emotional state; interactive teaching systems where an alert

to the teacher is displayed depending on the behavior of its student; virtual visit of museum with automatically displayed information based on the gaze of the user; medical research on behavior disorders; independently on the application the system should be also able to provide automatic assistance anticipating the user needs so as to obtain a proactive system. To summarize developing such multimodal human-machine interaction systems faces interesting applications issues as well as exciting research challenges.

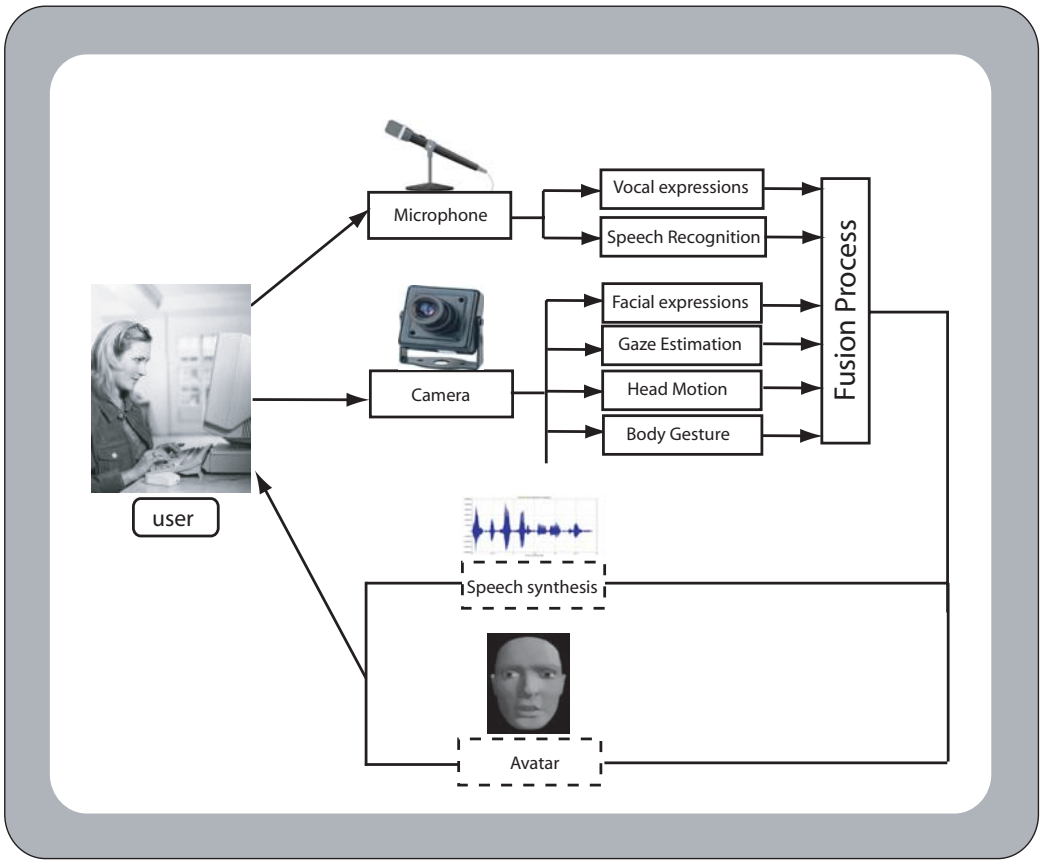

Figure 5.8: Multimodal Human-Computer interaction system .

Conclusion and Perspectives

Conclusion

Dans ce travail nous avons proposé différentes contributions pour la mise en place d'un système automatique de classification d'expressions faciales.

Pour la segmentation des traits du visage, nous avons développé des algorithmes pour extraire les contours de l'iris, des yeux et des sourcils basés sur la maximisation d'un flux de gradient de luminance autour des contours des traits recherchés. Les modèles proposés sont flexibles et permettent d'être robuste aux conditions d'éclairage, à l'origine ethnique, au port de lunettes et aux déformations de ces traits qui peuvent survenir sur le visage en cas d'expression faciale. Pour la segmentation des lèvres, nous avons utilisé un algorithme précédemment développé au laboratoire.

A partir de l'information portée par ces contours il a été montré qu'il est possible de reconnaître les expressions faciales. Cinq distances caractéristiques sont définies à partir des traits permanents du visage. Dans le but d'être le plus possible indépendant des variabilités entre les individus, ce sont les déformations par rapport à l'état *Neutre* qui sont utilisées pour la modélisation des expressions faciales à partir du Modèle de Croyance Transférable (MCT). L'utilisation de ce modèle a prouvé sa capacité à travailler avec des données imprécises obtenues à partir d'une segmentation automatique, à modéliser le doute entre plusieurs expressions et finalement à modéliser les expressions inconnues.

En plus de ces propriétés, nous avons proposé une extension de la méthode proposée pour la reconnaissance de séquences d'expressions en se basant sur l'introduction de l'information temporelle dans le MCT. La reconnaissance est basée sur la combinaison de l'ensemble des déformations des traits du visage entre le *début* et la *fin* de la séquence de l'expression. Cette amélioration permet d'être plus robuste aux erreurs ponctuelles de segmentation et aux déformations asynchrones des traits du visage.

Nous avons par ailleurs effectué une étude préliminaire sur la reconnaissance des expressions vocales dans le but de développer un système bimodal (audio + vidéo) de classification d'expressions. L'analyse du signal de parole est basée sur des paramètres prosodiques et la classification est faite par un SVM. Les résultats montrent que lors de l'analyse du signal audio, il faut plutôt considérer deux classes d'expressions vocales: *active* et *passive* plutôt que six classes comme c'est le cas pour la vidéo.

Perspectives

Étant donnée la complexité de la tâche de reconnaissance des émotions, nous proposons quelques pistes pour améliorer le système de classification d'expressions faciales proposé.

Dans ce travail nous avons constaté que certaines expressions sont difficiles à dissocier. Notre hypothèse était que les traits permanents du visage comportent suffisamment d'information pour reconnaître les 6 expressions universelles. Pour ce faire nous avons utilisé des points caractéristiques provenant des contours des traits du visage. Cependant la forme de ces contours peut aussi être une information importante à prendre en compte afin de mieux séparer certaines expressions confondues. Par exemple, les résultats ont montré que les distances caractéristiques ne sont pas suffisantes pour dissocier dans tous les cas le *Sourire* et le *Dégoût* alors que la forme des lèvres est très caractéristique pour chacune de ces expressions. Donc une première amélioration de notre modèle de reconnaissance est l'introduction de la forme des contours des traits du visage. L'introduction de cette nouvelle information dans notre processus de classification est rendue facile grâce à notre architecture de fusion basée sur le MCT. L'information de forme peut être vue comme un nouveau capteur avec son propre modèle de confiance.

La méthode proposée a été développée dans le cas restrictif d'expressions faciales sans aucun mouvement de lèvres. Cependant, la complémentarité entre les deux modalités audio et vidéo est évidente. La première étape est donc de quantifier l'influence sur les résultats de la classification des mouvements des lèvres dans le cas où le sujet parle en même temps qu'il est dans un état expressif (par exemple dans le cas où une personne sourit en même temps qu'elle dit "Je suis contente de vous voir"). Ce travail représente une première étape vers une combinaison des informations faciale et vocale pour un système bimodal de classification d'expressions.

Parmi les axes de recherches que nous avons choisis de suivre dans le travail présenté, nombreux sont basés sur l'état de l'art de la recherche en psychologie et aussi sur nos propres expériences. Cependant cette approche multidisciplinaire doit être étendue pour procurer des informations plus précises sur la capacité humaine à reconnaître les expressions faciales dans des conditions naturelles. Notamment plusieurs questions restent ouvertes: quelles sont les informations présentes dans le visage et utilisées par le système visuel humain pour reconnaître les expressions faciales? Existe-il un modèle de mouvement pour chaque trait du visage qui caractérise chaque expression? Quelle est l'information présente dans la parole et utilisée par le système acoustique humain pour reconnaître les expressions? Comment le cerveau humain fusionne les deux modalités et quelles sont les propriétés de cette combinaison (par exemple modèle de fusion linéaire/non linéaire) en fonction du contexte d'application?

Ce travail multidisciplinaire est d'un intérêt majeur pour notre modèle de reconnaissance d'expressions faciales. De plus, les réponses à toutes ces questions amèneront de nouvelles informations qu'il faudra également fusionner. Ceci pourra être facilement fait par notre modèle basé sur le MCT. Inversement notre modèle et l'étude de ses performances dans n'importe quelles conditions expérimentales donne un retour d'un intérêt majeur pour les psychologues qui veulent évaluer les limites de leurs modèles comportementaux.

Toutes ces questions vont être étudiées dans le cadre de mon post-doc au *Département de psychologie de l'Université de Montréal* dans l'équipe de recherche de Frédéric Gosselin.

Conclusion and Perspectives

Conclusion

In this work we have proposed several contributions for the realization of an automatic facial expressions classification system.

Concerning the facial features segmentation process, we have developed an algorithm for the segmentation of the iris, eyes and eyebrows based on the maximization of the luminance gradient around the contours of the required features. The proposed models are flexible and allow to be robust to luminance conditions, ethnicity, spectacles and facial features deformations during a facial expression. For the lips segmentation, we have used an algorithm previously developed in our laboratory.

Based on these facial features contours it has been shown that it is possible to classify facial expressions. Five characteristic distances are defined from the permanent facial features. So as to be independent of the variability between people, their motion according to the *Neutral* state is then used for the facial expressions modeling process based on the Transferable Belief Model (TBM). The use of this model has proved its ability to deal with imprecise data obtained from automatic segmentation, to model the doubt between some expressions and finally to model the unknown expressions.

In addition to these properties, we have proposed an extension of our method for the recognition of sequence of expressions based on the introduction of the temporal information in the TBM. The recognition is based on the combination of the whole set of facial features deformations between the *beginning* and the *end* of the expression sequence. This improvement allows to be more robust to punctual segmentation errors and to asynchronous deformations.

In order to extend the proposed system, we have realized a preliminary study on vocal expressions recognition towards a bimodal expressions classification system. The analysis of speech is based on prosodic parameters which were trained on a SVM for the recognition process. Results show that two main vocal expressions classes appear: *active* and *passive* expressions.

Perspectives

Due to the complexity of emotion recognition task we propose to improve our facial expressions classification system.

In this work we have seen that some expressions are difficult to dissociate. Our assumption has been that the permanent facial features contain enough information to recognize the 6 universal facial expressions. To do so we have used some characteristic points belonging to the contours of these facial features. In order to dissociate between some expressions, the shape of these contours may also be an important information to be taken into account. For example the results have shown that the characteristic distances are not sufficient to dissociate between *Smile* and *Disgust* rather than the mouth shape is very characteristic to dissociate between them. Then a first improvement to our facial expressions model is the introduction of the facial features shape. The introduction of this new information in our classification process is made easy thanks to our fusion architecture based on the TBM. Shape information can be seen as a new sensor with its own confidence model.

The proposed method has been developed in the restrictive case of facial expressions without any lips motion due to extra movement such as vocal pronunciation. However we found evidences of complementarity between the audio and visual modalities. Then the first step is to quantify the influence of the lips motion in the case when the subject is speaking while being in an expressive state (for example in the case of somebody smiling while telling "I am happy to see you"). This work represents a first step towards the combination of the face and voice information for a bimodal expressions classification system.

Many of the research directions we have chosen to follow in the presented work have relied on the study of the psychological literature and also on our own experiments. However this interdisciplinarity approach has to be extended to provide more precise information on the human ability to recognize facial expressions in many natural conditions, and, beyond, the emotions felt by somebody. Notably many questions remain opened: what is the information present in the face and used by the human visual system to recognize facial expressions? Does it exist a specific temporal motion model for each facial features which characterize the expressions? What is the information present in the speech and used by the human acoustical system to recognize vocal expressions? How does the human brain fuse these two modalities and what are the properties of this combination (for example linear/non linear fusion model) according to the context of the application?

This interdisciplinary work is of major interest for our model of facial expressions recognition. Again all the preceding questions deal with additional information to be integrated or fusion problems which are made easily tractable in our model based on the TBM. Inversely our model and the study of its performances in any experimental conditions provides a feedback of major interest for psychologists who aim at evaluating the strength and limits of their behavioral models.

All these questions will be studied in a future collaboration work in the scope of my post-doctoral position at the *Département de psychologie de l'Université de Montréal* in the research group of Frédéric Gosselin.

Part IV

Appendix

Chapter 6

Iris segmentation used for gaze direction and vigilance estimation

This work has been realized in collaboration with Corentin Massot, PhD student at the Laboratory of Images and Signals in Grenoble in France. In the following I present the result of our collaboration [Ham05g].

6.1 Introduction

During the past two decades a considerable scientific effort has been devoted to understand human vision. Since early works on the visual process study (see [Bus20]), many promising applications have considered eyes movements as well as vigilance characteristics as behavioral information (see [Zhu05], [Han05],[Nou05], [Wan05]), in order to develop sophisticated human-machine interfaces. In recent years, the evolution of the user interfaces for computer systems has been producing a significant impact since they are oriented for example to the development of intelligent multi-modal interfaces, gaze-contingent smart graphics or assisting systems for people with disability. The key idea for those systems is to make the communication/interaction with machines more intuitive.

Many of these real-world applications require accurate and real time iris segmentation and a lot of scientific effort has been dedicated to this field.

6.2 Existing systems

Tracking and recording a person's eye movements has been shown to be useful in diverse applications ([Lei83], [Mor99], [Yan99], [Tal99], [Sib01]). Notably eye gaze estimation and blink detection play an important role in human communication and are useful cues when attempting to understand user's behavior and intentions. They can be used as an interface between humans and computerized devices. For example they can enrich our way to interact with devices with seamless and more enjoyable mean of communication than keyboard and mouse. They can also give a mean of interpreting the user's behaviours like attention and vigilance for example in order to automatically detect an increasing risk of accident.

Eye gaze and blink estimation are two important applications of iris segmentation. Estimatimation technics are devided into two main approaches: intrusive systems (electrical skin potential and contact lenses) and non-intrusive systems (remote detection of the iris state). Here we present an eye gaze and vigilance estimation system based on the result of our iris segmentation technique which is non-intrusive (see Part I section 2.2.2). A lot of systems have previously been developed based on such an approach. These systems can be distinguished between IR based and the digital based systems.

The IR-based eye trackers (see Figure 6.1) generally exploit the center of the eye and the glint (reflection) which are easily obtained ([Ji 01],[Hut89],[Ebi93],[Mor99]). Most of the time assuming a static head, this methods use the glint as a reference point: the vector from the glint to the center of the pupil describes the gaze direction.

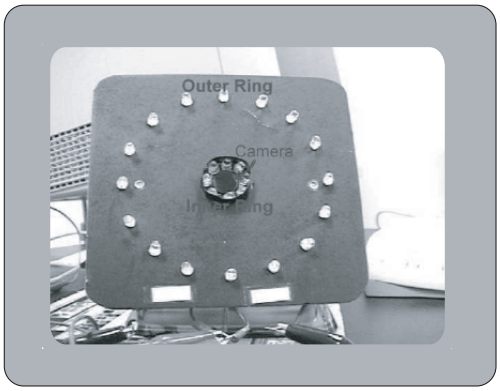

Figure 6.1: Example of infrared system [Qia03]

Tan *et al.* [Tan02] adopt a learning approach using a nearest neighbor and linear interpolation in an IR appearance-based method. They use locally linear embedding [Row00] of the view manifold. 252 samples images of 20x20 pixels are needed during the learning step to obtain an accuracy of 0.38^o using a leaving-one-out test.

Ji and Zhu [Ji 02] suggested a neural network-based method. They computed an eye gaze point by mapping pupil parameters to screen coordinates using generalized regression neural networks (GRNN). This system uses pupil parameters, pupil-glint displacement, ratio of the major to minor axes of the ellipse fitting the pupil, ellipse orientation, and glint image coordinates. No calibration is necessary and it allows some natural head movements. An average precision of 5^o is obtained.

Morimoto *et al.* [Mor99] used the mapping functions from a glint-pupil vector to a screen coordinate of gaze point. The functions are represented by two second order polynomial equations, and the coefficients of the equations are determined by a calibration procedure. These methods demonstrates enhanced performance when the user's head is in a fixed position. A precision rate of 3^o is obtained.

[Yoo05] proposes a gaze estimation system based on the projection of the four corners of the screen mediating the four corresponding glints on the iris. The analysis of the projection deformation of the rectangle formed by the glints and the detection of the center of the iris leads to the estimation of the gaze. A small calibration phase using the four corners is required to increase the precision up to 0.82^o. This system allows free head movement and relies on the hypothesis that the user is at a usual work distance of the screen (50-60cm).

The IR-systems simplify the detection of the iris but their use is restricted to this specific task. On the contrary the use of standard digital based systems appears to be a great challenge when considering the large set of applications that these systems can handle with (for example: facial expression recognition, head motion and body gesture estimation). It has to be stressed that for the two approaches, the segmentation accuracy mainly depends on the resolution of the camera.

[Wan05] models the iris contours as two planar circles and estimate projections onto a retinal plane. Using the ellipsoidal shape of projected eyes, anthropometric knowledge and the known distance to the subject for gaze determination they achieve a 0.5^o precision.

[Mat00] presents an eye gaze estimation method in which eye corners are located using a stereo vision setup and a 3D head model. The eyeball position is calculated from the pose of the head and a 3D offset vector from the mid-point of the corners of an eye to the center of the eye. A precision rate of 3^o is obtained.

[Han05] considers the transformation from image to screen coordinates via the eye as an homography. An homography is defined by 4 points and hence the transformation from image to screen coordinates is defined by at least 4 points. If the head moves together with the eyes when the gaze is shifted on the screen, additional calibration points are needed; thus, 4 points can only be considered as a lower bound. To estimate the gaze, they apply a simple calibration procedure relying on 4 points assuming that the head is kept fixed, that the eye is spherical and that the user is not too close to the screen plane. An average precision of 4^o is obtained.

In addition to gaze estimation, eyes movement is also one of the visual behavior cues that reflect a person's level of fatigue. As it was validated in [Din98], the ratio of eye closure over time is the most valid ocular parameter for monitoring fatigue.

One of the recent work in this field is [Ji 04] which uses the percentage of the eye closure over time (PERCLOS) and the average eye-closure speed (AECS). The program continuously tracks the person's pupil shape and monitors eye closure at each time analyses the ratio of pupil ellipse axes. PERCLOS and AECS are computed into a 30s window and display them onto the computer screen in real time, which allows an easy analysis of the alert state of the driver. These two parameters are then included into a complete set of cues (e.g. head movement, yawning frequency, gaze) and a bayesian network realizes the fusion between all of them in order to monitor and predict driver fatigue.

[Smi05] also uses several parameters to detect the vigilance level of the driver and among them the blinking frequency. The state of the eyes is determined evaluating the pixels intensity of the detected eye region in the current frame in comparison with its value in the first frame. If the eyes are closed for more than 40 out of the last 60 frames then the system warns that the driver has a low visual attention level. Specifically, the number of frames where the driver has his eyes closed are counted. Many studies have been done to determine the duration the eye must remain closed inside a time interval in order to detect a decrease of the driver's visual attention.

Infra-red cameras are very appropriate because they avoid any false detection of the iris and simplify the segmentation technique. However these cameras are expensive and they are only dedicated to iris detection. On the contrary digital cameras become standard device in most of common digital systems such as mobile phone, PDA and can be directly integrated in computer screen. They become cheaper while increasing their resolution. They can be applied in many applications such as visual communication, teleconferencing, remote environment visualization. But working with digital cameras is a difficult challenge because

the only available information are the pixels intensity of the image. This constraint introduces more complexity in the detection, the tracking and the segmentation process.

Here we present two applications of our iris detection technique based on digital cameras: a gaze estimation system and a vigilance estimation technique. These two systems can be integrated in a real human-machine environment. Their results also demonstrate the accuracy and robustness of the iris detection technique.

6.3 Gaze direction estimation

We present an approach to estimate the gaze direction of the user in front of a computer screen. The proposed method shows an accurate detection of the position of the iris in the face. It emerges as one promising alternative to the existing systems ([Eye05]), which usually require a device (IR cameras or two cameras) and/or impose acquisition and calibration constraints (helmet mounted on the head, great number of calibration points) that are awkward to use. Some of the additional advantages of our method are that it uses a commercially available video acquisition system made of a single camera (e.g., a webcam) placed above or below the screen (Figure.6.2). Our system makes the assumption that the head is kept fixed. This assumption can be removed using a head pose tracking system which is a problme that is out of the scope of our work.

6.3.1 Geometrical model and approximation of the projection function

We have to define the projection function, which establishes the relationship between the position of the iris center in an image and its corresponding position on the screen (projection). We define the following geometrical model (Figure.6.2 left). O is the center of the screen;

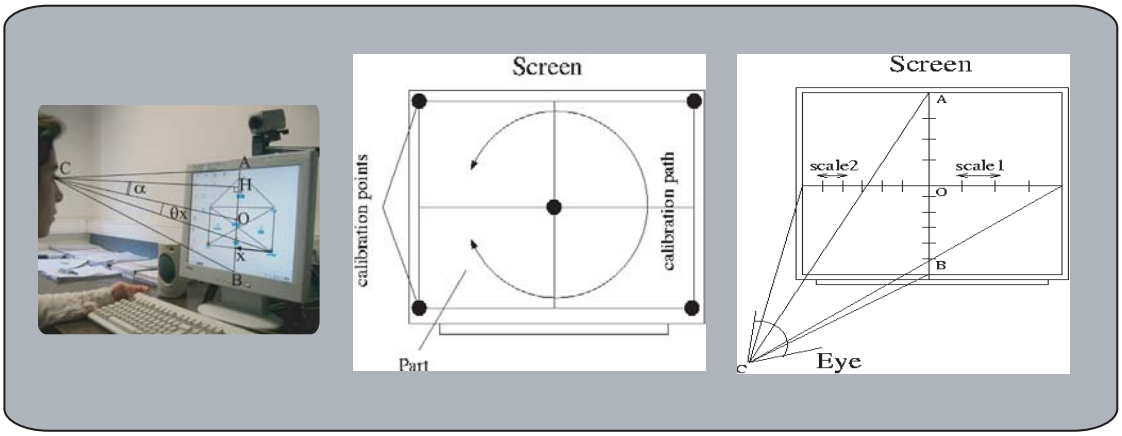

Figure 6.2: Left: overview of the system and geometrical model; middle: calibration configuration; right: scale configuration.

(A, B) represents the height of the screen (a similar model can be done for the width); C is the center of the iris; H is the orthogonal projection of C and is chosen as the origin of the screen plane reference; α is the angle between CH and CO and represents the position of the user relatively to the screen; x is the coordinate on AB of the point fixed by the user; the

angles between CO and Cx is noted θ_x, between CO and CB is noted θ_1 and between CO and CA is noted θ_2. In order to find the analytical formulation of the projection function, Hx is simply expressed by:

$$Hx = CH * tan(\alpha + \theta_x) \tag{6.1}$$

Equation 6.1 shows that the estimation of gaze direction depends on the distance to the screen CH and on the angle α. Figure.6.3 left presents the variation of the distance Hx with a fixed α and different values of CH. The analysis of these curves points out the fact that for $CH >= CH_{min}$ the projection function tends to be linear. Figure.6.3 middle presents the variation of the distance Hx with a fixed CH and different values of α. The analysis of these curves points out the fact that for $\alpha <= \alpha_{max}$ the projection function tends also to be linear. Figure.6.3 right presents the comparison between the projection function and its equivalent linear approximation. We finaly estimate that taking $CH_{min} = 3.OA$, i.e., $\approx 30cm$ and $|\alpha| <= |\alpha_{max}| = 10^o$, it is possible to consider the projection function as linear. These values correspond to usual work conditions justifying the use of a linear approximation for the projection function (Figure.6.3 right).

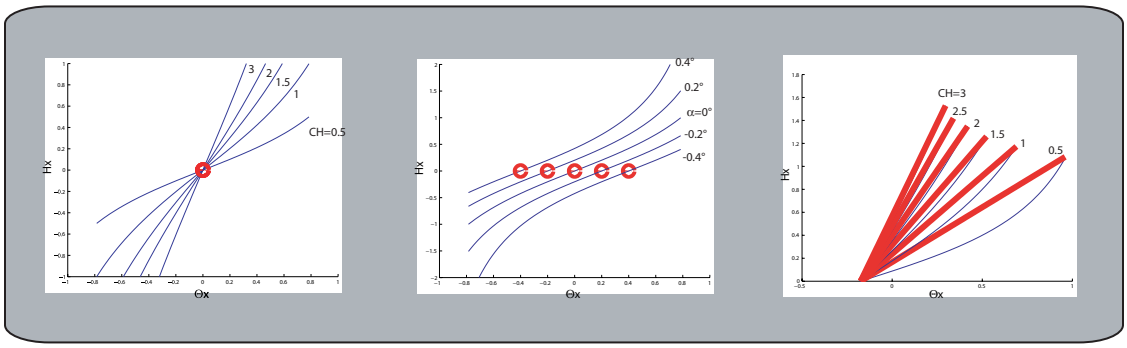

Figure 6.3: Evolution of the projection function in relation to CH (left), in relation to α and linear approximation (positive part of the function) (right).

6.3.2 Implementation

The screen is divided into four parts around the screen center O (Figure.6.2 middle). A specific scale is associated to each part. The step scale corresponds to the distance covered on the screen for one unit of movement of the iris (e.g.$1unit = 1pixel$). For example, scale 1 of Figure 6.2 right has to be greater than scale 2, because a displacement of one pixel of the iris center in the image plane will be greater in the top right part of the screen than in the top left part. Ensuring differences on the scales of each part of the screen at each coordinate allows a better approximation of the projection function. Indeed CO is not necessarily perpendicular to the screen (Figure.6.2 right). The condition $\alpha \leq \alpha_{max} = 10^o$ ensures a good linear approximation for the axis perpendicular to the axis of the camera (i.e if the camera is placed below the screen, the horizontal axis is well approximated). But in experimental conditions the linear hypothesis tends to reach the limit cases on the axis of the camera (i.e the vertical axis on our example). Allowing the system to define two different scales on this axis increases the

precision of the detection on this direction, removing this slight shift of the projection function (Figure. 6.3 middle).

We calculate the gaze direction using the projection of the iris center position on the screen. The system just takes into account the spatial displacement of the iris and not any velocity information. The process is divided into two steps : the calibration which automatically initializes the system and the detection itself.

a) **Calibration.** The calibration consists in automatically defining the x and y scales for each part of the screen. For this, five points are necessary : the center and the four corners of the screen (Figure.6.2 middle). In the calibration stage, the user has to be in front of the screen at a normal distance (50cm to 80cm), while a camera is located above or under the screen. When the calibration procedure begins, five points appear dynamically on the screen one by one beginning by the center and finishing at each corner. The user has to follow them with the eyes. The major difficulty is that human cannot precisely fix a given position without any dispersion. In order to overcome this problem, each point of calibration appear on the screen during 1 second. The system saves all the iris positions and compare them with the position of the iris when the user fixed the center of the screen (first calibration point). This position is taken as the origin of the iris plane reference. From all the positions of the iris center, four histograms are computed: two corresponding to the abscissas positions, the first one for the negative ordinates (relatively to the origin of the iris plane reference) and the second one for the positive ordinates; the two remaining histograms for the ordinates positions are defined in the same way. Then on each histogram, the two positions (on each side of the origin according to the abscissa axis or to the ordinates axis) that have been fixed the most often are chosen. After analysing all the histograms four couple of coordinates are obtained corresponding to the iris center position while fixing the four corners of the screen. These coordinates are used to define the four different scales of the four parts of the screen as described in 6.3.1. The dynamic appearance of the calibration points on the screen is implemented via the use of a free Matlab toolbox: the Psychtoolbox (http://www.psychtoolbox.org/), dvelopped by Brainard [Bra97] and Pelli [Pel97]. This one allows to very accurately define the position, the size and the exact time of appearance of the points. At each trial, the five calibration points first appear one by one; the calibration is performed just after the removal of the last point; then the target image is presented and the user can freely scan it or following a predefined trajectory according to the task.

b) **Detection of gaze direction.** For each image, the right and the left iris center positions are extracted. Each center is related to one point of the screen by the projection process. To eliminate transitory positions (occurring during the movement of the eye toward its new position), we introduce the notion of fixation which is a region fixed during several frames. According to the experimental data in the medical litterature [Cha95], a fixation takes between 200ms and 400ms. With 6 consecutive frames at 25 frames/second we obtain a fixation duration of 240ms which agrees with the real data. At the end of the sequence we obtain what we call a "fixation map", which represents the whole set of fixations. The fixations can overlap, so we associate at each position a value corresponding to the number of times that this position has been fixed. The gaze position of the user on the screen corresponds to the barycenter of the set of points that

belong to the same fixation region obtained in the fixation map. At each recorded gaze position is associated an error window which size corresponds to two units of the scales associated to the part of the screen at which this position belongs to. Both a spatial and temporal labeling of the region into connected components is realized taking into account the possible overlapping between regions. A region is considered as a fixation region if the number of fixation points belonging to this region reach the required amount (6 frames). This parameter allows to easily select between varying duration of fixations. All the process is done off-line, the number of fixation regions and their chronological order is automatically determined and the entire scan-path can be recovered. The final result consists in the whole set of barycenters of the fixation regions and the temporal order in which they appeared during the sequence (a number is associated) (Figure.6.4 top, Figure.6.5).

6.3.3 Precision measures to determine the system performance

In order to evaluate the precision of our gaze direction estimation, an experimental setup, consisting in a grid of 18 black points plotted on the screen (Figure.6.4 bottom) is used to estimate the user gaze position during the fixation of these points. The camera is placed under the screen (1024x768), considering that the usual work conditions in front of a computer screen, vary in a range between 50cm to 80cm. The subject was asked to sequentially fix the black points and this experiment is carried out 10 times. After the experiment the mean euclidean error between each point and their associated fixation is computed and is converted into an error angle in order to remove the dependence to the distance from the screen variable. We obtain a mean precision of 0.8^o which means that the fixations are accurately detected. Figure.6.4 top presents the result of the estimation of the user gaze direction on the grid defined before. The white circles represent the estimated user mean fixation position for each fixed black point. These results are precise enough to realize different tasks such as pointing icones or explorating displayed images.

One of the potential application is the study of the strategy of a document exploration, like in figure 6.4 bottom which presents the analysis of a geographical map. The system is able to detect the user attention during the exploration. Each point represents the position of the gaze position on the map and its temporal apparition.

6.3.4 Comparison with a commercial system

We present two kinds of comparison results: a fixation map and a trajectory map. In the fixation map, the user has to fix each icon presented in the image in a free order. On Figure 6.5 left, each point represents one fixation (barycenter of a region of fixation). The associated number indicates the order in which the different icons have been looked at by the user. The video sequences are acquired by a camera with a frame rate of 25 frames per second. They are acquired when the user is in a position corresponding to usual work conditions in front of a computer screen with the constraint that the head is kept fixed.

In order to evaluate the performances of our detection system, we have carried out the same experiments in exactly the same conditions with a commercial Eye-tracker: the SR Research-EyeLink system ([Eye05])(Figure.6.5 right). This system is made of two infra-red cameras mounted on an helmet set which gives an extremely good definition of the iris. The system is very accurate (precision $< 0.5^o$) in the best conditions.

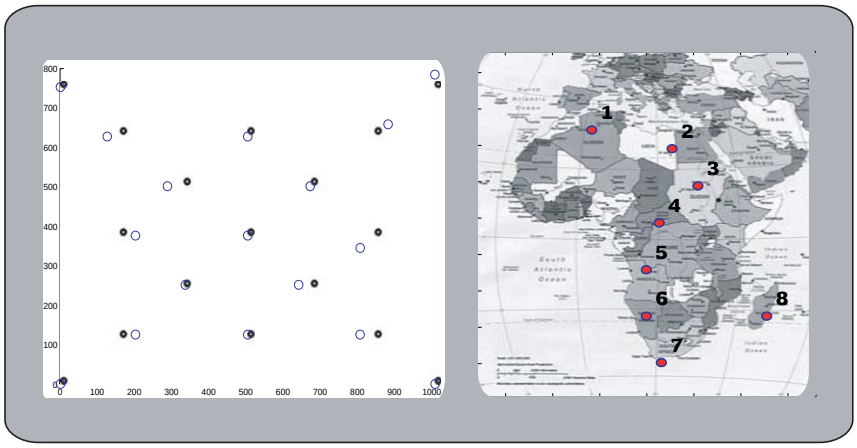

Figure 6.4: Top : grid made of black points corresponding to the different positions used to estimate the precision (the size is 1024x768 pixels and corresponds to the whole screen); white circles represent the results of the user gaze detection on the grid; Bottom : analysis of the exploration strategy of a geographical map; the points are situated on the really observed countries with their chronological order.

In Figure.6.5 and 6.6 we aim at rebuilding the ocular trajectory of a user during two different tasks. In figure 6.5 the task was to successively fix predefined icons. Our results (Figure.6.6 left) have been compared with those obtained with the Eye-Link (Figure.6.6 right) in the same experimental conditions (luminance, position of the observer, same duration). The number of fixations is not exactly the same because the Eye-link has an on-line process and a fixation is defined as a position between two saccades. A saccade is detected measuring the speed and acceleration of the change of the gaze position. On the contrary our system detects a fixation if the spatial region contains a constant amount of gaze positions (6 frames). However the fixations points appear very similar both in position and systematic error. In Figure.6.6

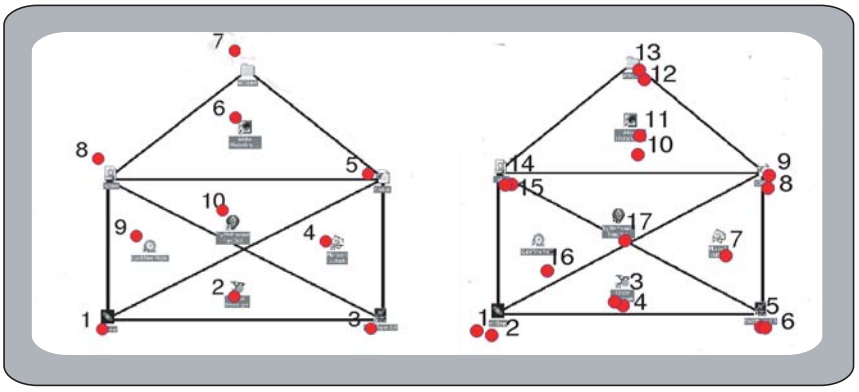

Figure 6.5: Results of icons fixation. Left: fixation map with our system; right: fixation map with Eye-Link.

the task was to continuously follow the edges of a drawn house. Our results (Figure.6.6 left)

have been compared with those obtained with the Eye-Link (Figure.6.6 right) in the same experimental conditions. The Eye-link trajectory is composed of all the detected fixations points which are then connected. So the global trajectory shape appears rectilinear. On the contrary, our trajectory is composed of the whole set of points corresponding to the gaze locations. A final weak smoothing removes some perturbations due to the difficulty for the observer to keep permanently the gaze on the edge. This is the reason why the trajectory shape appears more agitated than the Eye-link one. However our trajectory reflects better the real scan path followed by the observer during the task. Again the results obtained with the two systems appear quite similar. The comparison of the detection obtained with both

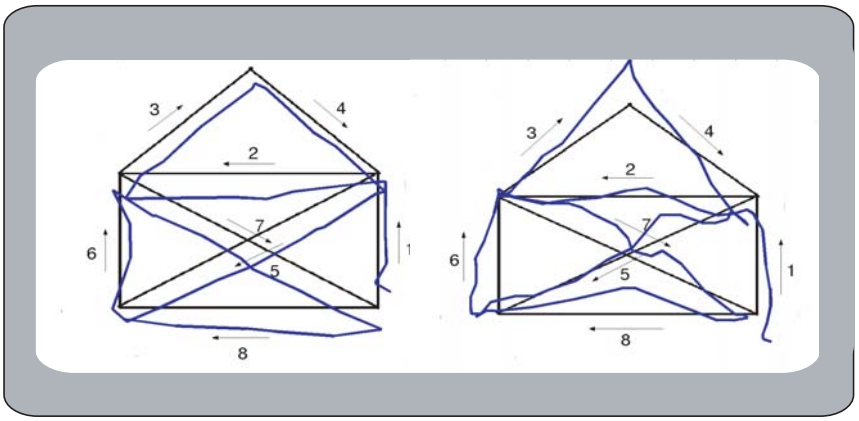

Figure 6.6: Left: trajectory map with our system; right: trajectory map with Eye-Link.

systems points out a comparable quality of our results in the specific tasks described before. For the two considered tasks our system gives good results and performs gaze detection with sufficient accuracy to be incorporated into real applications with the only constraint that the head is kept fixed.

6.4 Hypovigilance estimation

The estimation of hypovigilence of a user (related to his fatigue level) can be performed by the evaluation of the blinking frequency (see section2.2.2.1). In case of "normal" vigilance level, the medical literature gives values between 12 to 20 per minute [Bli02]. Deviations from this frequency can then be interpretated as a change in the level of vigilance of the user. To automatically detect an hypovigilence, first the state of the eye (open or closed) has to be detected. This is done by detecting the blink of the eyes. Then integrating the blinks over a temporal window allows to recover the blink frequency which is finally analysed.

To analyze the vigilance, different sequences have been acquired with a standard camera at 25 frames per seconds and during 10 minutes. The subjects were asked to simulate three states: normal blinking, fast blinking and drowsiness respectively. The blinking frequency evaluated on our sequences is in average 18 blinks per minute. To estimate the level of vigilance, the blink frequency is computed at each time t inside a temporal window Δ_t analyzing the last 6 seconds (period of one blink at a normal blinking frequency). The detection of one or two blinks corresponds to a normal blinking frequency. If the frequency is higher, the ratio between the duration of the open eye states and the closed eye states indicates whether it is

a case of fast blinking (eyes are found open two times longer) or a case of drowsiness (eyes are found closed two times longer). If the frequency is lower, then the same ratio indicates whether it is a case of normal blinking (the eyes are kept open during all the temporal window) or a case of drowsiness (the eyes are kept closed).

Figure.6.7 shows four frames taken from one of the sequences described below and examples of detection of the eyes state and of the vigilance level.

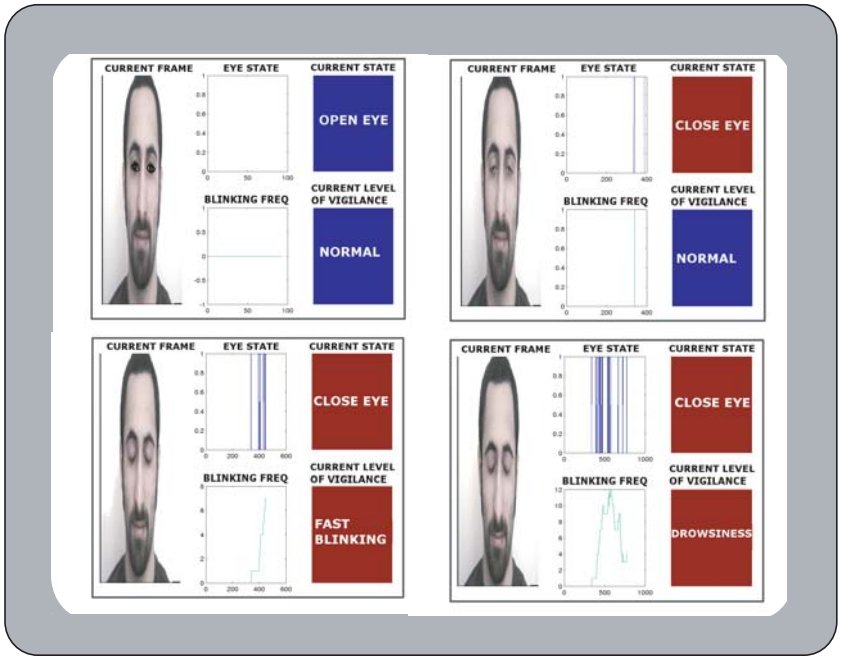

Figure 6.7: Estimation of the vigilance. For each figure: left: current image; middle, top: past evolution of the eyes states, bottom: past evolution of the blinking frequency; right, top: current detected eye state, bottom: current vigilance level.

6.5 Conclusion

Based on the iris segmentation results we have presented two associated applications in human-machine interaction domain: an accurate and low constrained gaze direction estimation system working with a single commercially available video acquisition system. Secondly, we presented an analysis of the frequency of the blink, which can be used in the detection of the vigilance level of the user. Preliminary results and comparative experiments with existing systems show the interest and the robustness of the proposed approach.

Chapter 7

Facial expressions classification based on the Bayesian model and on HMMs

This work has been realized in the scope of a PhD exchange program in the Network of Excellence SIMILAR [Sim04] in collaboration with the laboratory of *Théorie des Circuits et Traitement du Signal* (TCTS) in Mons in Belgium, in the research group of Professor Thierry Dutoit. I had the opportunity to meet Laurent Couvreur a statistician whose speciality is speech recognition and classification. For research curiosity, we have tested our data used for facial expressions classification on his implementation of Bayesian model and HMMs. The modeling phase were efficient however quickly we had to face a lack of data for these two classifiers. In the following I present the result of our collaboration [Ham05e].

7.1 Facial expressions classification based on the Bayesian model

7.1.1 Method

In order to demonstrate the effectiveness of the approach based on the Transferable Belief Model (TBM) for the purpose of facial expressions classification, we compare it with other standard classification techniques. More especially, we apply the well-known Bayesian model [Dev82]. The classification problem can be formulated in terms of a Bayesian classification problem. Consider the set $\Omega_E = \{E_1, E_2, ...E_8\}$ of all the considered facial expressions including the *Unknown* expression E_8. Unlike in the TBM, the Bayesian model considers only singleton hypotheses in the classification problem. Besides, the *Unknown* expression is viewed as an extra hypothesis that the Bayesian theory will model explicitly, contrary to the TBM that does not really model this hypothesis but rather derives its belief from the other modeled hypotheses. During the modeling step, one can model the a posteriori probability of a facial expression E_i given some observed characteristic distances D_i. During the classification step, the best hypothesis $E^*_{D_i}$, i.e. the most probable facial expression, with respect to the characteristic distances D_i is found by maximizing the a posteriori probability:

$$E^*_{D_i} = arg(\max_{E \in \Omega_E}(P_{D_i}(E|D_i))) \tag{7.1}$$

Obviously, conflicts can arise if the probability models of two characteristic distances lead to different decisions. These conflicts can be solved in various ways. For example, one can use voting techniques to select an hypothesis among the ones proposed with respect to different characteristic distances. Alternatively, the decision can be based on all the characteristic distances simultaneously and made by relying on the joint a posteriori probability $P_{D_i}(E|\overline{D})$ where $\overline{D}$ denotes the observation vector $[D_1, ..., D_5]^T$ gathering together the five characteristic distances. In this case, the decision rule becomes:

$$E^* = arg(\max_{E \in \Omega_E}(P(E|\overline{D}))) \tag{7.2}$$

Using the Bayes rule [Dev82], the decision rule can be written as:

$$E^* = arg(\max_{E \in \Omega_E}(\frac{P(\overline{D}|E)P(E)}{P(\overline{D})})) \tag{7.3}$$

where $P(\overline{D}|E)$ is the joint posterior probability, $P(E)$, the *a priori* probability on the expressions and $P(\overline{D})$ the probability associated to the observation vector.

Here, we assume that all the facial expressions have equal a priori probabilities [1]. Besides, the data probability at the denominator of the equation 7.3 is constant for each hypothesis. Therefore, the decision rule reduces to a maximum likelihood classification problem:

$$E^* = arg(\max_{E \in \Omega}(P(\overline{D}|E))) \tag{7.4}$$

The likelihood function for each facial expression is estimated during the modeling phase. It is current to assume a parametric model for these functions and to derive their parameters from observation data. Here, we consider Gaussian Mixture Models (GMM) [Bil98]. The likelihood function is represented as a weighted sum of Gaussian probability density functions, that is,

$$P(\overline{D}|E) = \sum_{k=1}^{K} w_{E,k} N(\overline{D}; \mu_{E,k}, \Sigma_{E,k}) \tag{7.5}$$

$$= \sum_{k=1}^{K} w_{E,k} \frac{1}{(\sqrt{2\pi})^d |\Sigma_{E,k}|^{1/2}} exp\left[-\frac{1}{2}(\overline{D} - \mu_{E,k})^T (\Sigma_{E,k})^{-1} (\overline{D} - \mu_{E,k})\right] \tag{7.6}$$

where $w_{E,k}$, $\mu_{E,k}$ and $\Sigma_{E,k}$ stand for the mixing coefficient, the mean vector and the covariance matrix, respectively, of the k-th component of the Gaussian mixture model for a given expression E_i. By definition, the coefficient d is equal to the dimension of the observation vector $\overline{D}$, namely $d = 5$. The number K of components is chosen as a tradeoff between model complexity and data availability. Clearly, the more components there are, the finer the model is. However, the number of parameters increases with the number of components and more data are required to estimate them accurately. In our experiments, $K = 3$ was shown to

[1] Remark that the equi-probability hypotheses is very strong while in the TBM is not necessary to make such a hypotheses on the data.

give the best results. The estimation of the GMM parameters can be cast as a Maximum-Likelihood estimation problem with incomplete data. Indeed, the problem is trivial if one can assign every observation vector to a Gaussian component of the mixture. For example, given a sequence $\{\overline{D_1}, ..., \overline{D_N}\}$ of N observation vectors (in our case $N = 5$), the estimate of the mean parameter $\mu_{E,k}$ of the k-th component is computed as follows:

$$\mu_{E,k} = \frac{\sum_{n=1}^{N} I_k(\overline{D_n})\overline{D_n}}{\sum_{n=1}^{N} I_k(\overline{D_n})} \tag{7.7}$$

where $I_k(\overline{D})$ is an indicator function that returns 1 if the observation vector actually comes from the k-th GMM component and returns 0 otherwise. Similar functions can be given for the other parameters. Unfortunately, such assignment is unknown. Hence, the estimation is performed via the Expectation-Maximization (EM) algorithm. The presentation of this algorithm is beyond the scope of this paper and we just overview the basic ideas. More information can be found in [Bil98], [Rab89]. The EM algorithm consists in an iterative procedure. It starts with initial values of the GMM parameters. Then, during the so-called Expectation step, it computes the a posteriori probabilities of the observation vectors for each GMM component. For example, given an observation vector $\overline{D}$, the a posteriori probability$P_k(\overline{D}|E)$ for the k-th component is:

$$P_k(\overline{D}|E) = \frac{w_{E,k} N(\overline{D}; \mu_{E,k}, \sum_{E,k})}{\sum_{k=1}^{N} w_{E,k} N(\overline{D}; \mu_{E,k}, \sum_{E,k})} \tag{7.8}$$

where $w_{E,k}$, $\mu_{E,k}$ and $\sum_{E,k}$ stand for the mixing coefficients, the mean vector and the covariance matrix, respectively, of the k-th component of the Gaussian mixture model for a given expression E. The number K of components is chosen as a tradeoff between model complexity and data availability. However, the number of parameters increases with the number of components and more data are required to estimate them accurately. In our experiments $K = 3$ was shown to give the best results.

It allows distributing *softly* the observation vector among all the GMM components, the distribution being weighted by the a posteriori probabilities. Next, new estimates of the GMM parameters can be computed as in the trivial case where each observation would be assigned "hardly" to a single component, except that weighted terms are involved in the estimation formula. Let us come back to our example of computing the mean parameter $\mu_{E,k}$. During the so-called Maximization step, the EM algorithm computes a new estimate of this parameter by replacing the indicator function with the a posteriori probability,

$$\mu_{E,k} = \frac{\sum_{n=1}^{N} P_k(\overline{D_n}|E)\overline{D_n}}{\sum_{n=1}^{N} P_k(\overline{D_n}|E)} \tag{7.9}$$

Similar re-estimation formula are defined for the other parameters. The two steps of the EM algorithm are repeated alternatively until convergence is reached. Practically, the

System \ Expert	E_1	E_2	E_3	E_7	E_8
E_1 *Smile*	**37.71**	3.80	21.86	5.46	23.96
E_2 *Surprise*	22.27	**50.43**	3.79	4.94	19.17
E_3 *Disgust*	4.33	10.16	**25.43**	5.20	12.79
E_7 *Neutral*	7.62	20.47	2.21	**79.85**	24.56
E_8 *Unknown*	28.07	15.14	46.71	4.55	**19.52**

Table 7.1: Bayesian classification rates in percentage for the HCE database.

algorithm stops when the estimates of the parameters do not change significantly any more. The major problem of this estimation procedure is the initial conditions. A standard approach to find initial values of the parameters consists in clustering the observation vectors into as many classes as GMM components. Many clustering algorithms can be used and we adopted a K-means algorithm [Lin80] in this work. Once the observation vectors have been grouped into clusters, initial estimates of the parameters of the GMM components can be estimated as in the trivial case.

7.1.2 Results on Bayesian model

For the Bayesian classification system, all the data of HCE database are used in the training procedure. The data correspond to the characteristic distances obtained on manual segmentation to be independent of the segmentation errors. The test is carried out by a $21 - fold$ cross validation to be consistent with other Bayesian based classification approaches [Coh03a]. It consists in taking 20 out of 21 subjects for training and the remaining subject for test step. The process is repeated 21 times, considering a different test subject each time. The classification rate is the average over 21 results.

Unlike for the TBM system, the *Unknown* expression has to be modeled as any other expression and requires training material. We actually assigned to the *Unknown* class all facial expression which does not corresponds to *Smile*, *Fear*, *Disgust* and *Neutral* expressions. Classification rates of the Bayesian classifier are given in Table 7.1.

The best result is for the *Neutral* expression and low rates are observed for the other expressions. These poor performances may be due to the fact that the assumptions (e.g., parametric model of the statistical distributions) that the Bayesian model relies on are questionable in our application. GMM are very efficient models when the number of components is large enough and enough material is available to estimate the mixture parameters. Here, we only tested one approach (fixed mixture size, ML estimation with full EM algorithm). However, there exist several variants (various initialization schemes for full EM algorithm, greedy EM algorithm Figueiredo-Jain (FJ) algorithm, non-linear least square estimation by Gauss-Newton (GN) algorithm, Markov Chain Monte Carlo (MCMC) method, Cross-Entropy (CE) algorithm). Even if we have a better classification rates with these more or less complex techniques we can not overcome the main limitations of the Bayesian model for facial expressions classification which mainly consists in the modeling of doubt between several expressions and of the *Unknown* expression. Indeed, contrary to the TBM in the Bayesian model the *Unknown* expression corresponds to a new class that represents a finite set of expressions added to the already defined ones. It does not contain all the possible facial configurations which

can lead to classification errors in the case of intermediate states between expressions, new facial expression or a sensors errors.

Moreover human is not binary and a doubt between some expressions can appear. Contrary to the TBM the doubt state does not exist in the Bayesian modeling. Based on these observations we can conclude that based on our data the Bayesian model is less adapted to model facial expressions than the TBM which are suitable to model the knowledge.

7.2 Facial expressions classification based on the HMMs

7.2.1 Method

In order to demonstrate the effectiveness of the classification based on dynamic data, we propose another decision process that takes into account the time dependency between neighboring frames in order to improve the recognition performance.

Here, we assume again that the current facial expression is mainly dependent of the facial expression at the previous time. Consequently, the time sequence of facial expressions together with the corresponding sequence of characteristic distance vectors can be viewed as the realization of a Hidden Markov Model (HMM) [Rab89]. From a generative perspective, a HMM selects a state (here, a facial expression) and produces an observation (here, a vector of characteristic distances) at each time. In the HMM framework, the state sequence is assumed to be a first-order Markov chain (at each time t we take into account the information present at time $t-1$), and the observations to be probabilistic functions of the states. The model is said *hidden* because only the observations are available while the underlying state sequence is kept unrevealed. In order to define a HMM, we first characterize the topology of the Markov chain. It is classical to display it as a transition graph where the nodes represent the states. In our case, the states are related to the facial expressions. Without loss of generality, we are only interested in modeling video sequences where the subject starts in the neutral state, performs a given facial expression and returns to the neutral state. The HMM topology has to depict such behavior, hence only some transitions are authorized as shown in Figure 7.1. To complete the characterization of the Markov chain, we define the state transition probabilities a_{ij} from the state s_{t-1} at time $t-1$ to the state s_t at time t and the initial state probabilities π_i as:

$$a_{ij} = P(s_t = E_i | s_{t-1} = E_j) \tag{7.10}$$

$$\pi_i = P(s_0 = E_i) \tag{7.11}$$

$$1 \leq i, j \leq 8$$

Given the topology of the Markov chain involved in the HMM for modeling facial expression sequences, most transition probabilities are constrained to be null and only the initial probability for the neutral state is not equal to zero.

Next, we need to define the state-conditional probability distributions, i.e. the statistical distribution functions of the random vector of characteristic distances given each facial expression state. Many models can be proposed to represent these distributions. As in the Bayes approach, we propose to represent every state-conditional distribution by its probability density function,

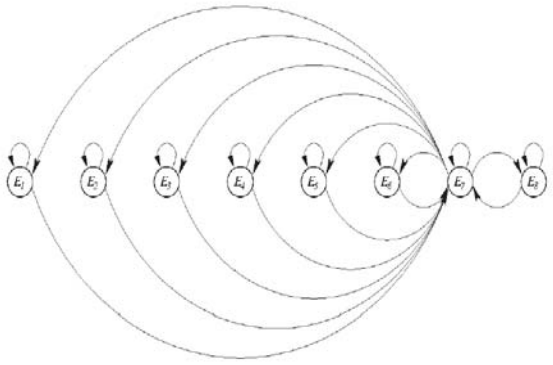

Figure 7.1: Topology of the HMM for modeling video sequences of facial expressions.

$$f_i(\overline{D}) = p(\overline{D}|s_t = E_i) \tag{7.12}$$

$$1 \leq i \leq 8 \tag{7.13}$$

with a Gaussian Mixture Model, that is,

$$f_i(\overline{D}) = \sum_{k=1}^{K} w_{i,k} N(\overline{D}; \mu_{i,k}, \Sigma_{i,k}) \tag{7.14}$$

As previously, the number of mixture components is set to $K = 3$. A major issue with HMMs is to estimate the model parameters $\Pi = [\pi_i]$, $A = [a_{i,j}]$ and $B = [f_i]$. Such estimation, which is classically performed in a Maximum Likelihood sense, is not a trivial problem since the data are doubly incomplete. Only the observation data (the characteristic distance vectors) are given and neither the underlying states (the actual facial expressions) nor the active GMM components are known. The problem is elegantly solved by the Baum-Welch algorithm ([Bil98], [Rab89]) that is actually an implementation of the EM algorithm for estimating Gaussian HMM parameters. It consists in a two-step iterative procedure that refines the estimate of the parameters. The initial values of the parameters can be obtained from an initial segmentation of the observation data, i.e. a labeling of the characteristic distance vectors with respect to the facial expression states. Another classifier can provide this first segmentation, for example the Bayesian classifier described in the previous section. Once the HMM parameters have been estimated, it can be embedded in a decision process to assign any given characteristic distance vectors to facial expression states. We consider the set $\Omega_E = \{E_1, E_2, ...E_8\}$ of all the considered facial expressions including the *Unknown* expression E_8. The approach consists in finding the best state sequence $S^* = \{s_1, s_2, ..., s_N\}$ given a N-length sequence of characteristic distance vectors. This search can be performed via the Viterbi algorithm ([Bil98], [Rab89]), a dynamic programming algorithm that finds efficiently the state sequence maximizing the joint likelihood $P(\overline{D_1}, ..., \overline{D_N}, S|\Pi, A, B)$, that is,

$$S^* = \underset{S \in \Omega_E}{argmax} P(\overline{D_1}, ..., \overline{D_N}) \tag{7.15}$$

Hence, a state is eventually assigned to every characteristic distance vector; equivalently a facial expression is assigned to every video frame. Note that the HMM approach can be viewed as an extension of the Bayesian approach. Indeed, the latter classifier takes a decision frame by frame without any constraint on the resulting sequence of decisions: all the decision sequences have the same a priori probability. Besides, the HMM classifier gives more or less importance to a decision sequence depending on the parameters of its Markov chain: the decision sequences can have different a priori probabilities. This property allows the HMM identifying the decision sequence globally and not as a concatenation of independent local decision.

7.2.2 Results on HMMs

For the HMM classification system, all the data of HCE database are used in the training procedure. The data correspond to the characteristic distances obtained on manual segmentation to be independent of the segmentation errors. The test is carried out by a $21 - fold$ cross validation. It consists in taking 20 out of 21 subjects for training and the remaining subject for test step. The process is repeated 21 times, considering a different test subject each time. The classification rate is the average over 21 results.

Unlike for the TBM system, the *Unknown* expression has to be modeled as any other expression and requires training material. It shares the same limitations as the *Unknown* class added in the Bayesian classifier (see Section 7.1). We actually assigned to the *Unknown* class all facial expression which does not corresponds to *Smile*, *Surprise*, *Disgust* and *Neutral* expressions.

Results are reported in Table 7.2. Similarly to our Frame-by-Frame classification, *Disgust* expression is lower than the others.

System \ Expert	E_1	E_2	E_3	E_7	E_8
E_1 *Smile*	**78.87**	0.0	2.02	3.28	14.45
E_2 *Surprise*	0.0	**79.81**	0.0	6.36	21.31
E_3 *Disgust*	6.82	0.0	**49.39**	3.60	28.66
E_7 *Neutral*	4.48	10.75	6.37	**75.25**	25.84
E_8 *Unknown*	9.83	9.44	42.22	11.51	**9.74**

Table 7.2: Classification rates of the system based on the HMMs for the HCE database.

As for the Bayesian model we can have better classification rates with more or less complex HMM modeling topology. However, we can not overcome the main limitations of this model for facial expressions classification which mainly consists in the modeling of doubt between several expressions and of the *Unknown* expression, as it is the case for the Bayesian model. Based on these observations we can conclude that based on our data the HMM based approach is less adapted to model facial expressions than the TBM which are suitable to model the knowledge.

7.3 Comment

This collaboration has been very fruitful. Indeed we had the opportunity to use other classifiers than the TBM. It allowed us to conclude that each approach has its own advantages and the choice of a specific classifier depends on the considered application.

Bibliography

[Abb04a] Abboud B. and Davoine F. Appearance factorization for facial expression analysis. *Proc. BMVA British Machine Vision Conference*, September Kingston, UK, 2004.

[Abb04b] Abboud B., Davoine F. and Dang M. Facial expression recognition and synthesis based on appearance model. *Signal Processing: Image Communication*, 19 no. 8:723–740, September 2004.

[Ami90] Amini A., Weymouth T. and Jain R. Using dynamic programming for solving variational problems in vision. *IEEE PAMI*, 12 no.9:855–867, September, 1990.

[And06] Anderson K. and W.McOwn P. A real-time automated system for the recognition of human facial expressions. *IEEE Trans. ON SYSTEMS, MAN, AND CYBERNETICS PART B: CYBERNETICS*, 36 no. 1,:96–105, February 2006.

[Ban96] Banse R. and Scherer K.R. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70 no.3:614–636, 1996.

[Bas78] Bassili J. N. Facial motion in the perception of faces and of emotional expression. *Experimental Psychology - Human Perception and Performance*, 4 no.3:373–379, 1978.

[Bas79] Bassili J.N. Emotion recognition: The role of facial movements and the relative importance of upper and lower areas of the face. *Journal of personality and social psychology*, 61 no. 11:298–307, 1979.

[Bea94] Beaudot W. *The neural information in the vertebra retina : a melting pot of ideas for artificial vision.* PhD thesis, TIRF laboratory, Grenoble, France, 1994.

[Bil98] Bilmes J.A. A gentle tutorial of the *em* algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical Report ICSI-TR-97-021, International Computer Science Institute, April 1998.

[Bla93] Black M.J and Anandan P. A framework for the robust estimation of optical flow. *Proc. Computer Vision, ICCV*, pages 231–236, Berlin, Germany, 1993.

[Bla97] Black M.J. and Yacoob Y. Recognizing facial expression in image sequences using local parametrized models of image motion. *Trans. Computer Vision*, 25 no.1:23–48, 1997.

[Bli02] Blink frequency. http://www.inrs.fr/inrs-pub/inrs01.nsf/intranetobject-accesparreference/tc+88/$file/tc88.pdf. 2002.

[Boe01] Boersman P. and Weenink D. Praat speech processing software. *Intitute of Phonetics Sciences of the University of Amsterdam. http://www.praat.org*, 2001.

[Bot02] Botino A. Real time head and facial features tracking from uncalibrated monocular views. *Proc. 5th Asian Conference on Computer Vision ACCV*, pages 23–25, January Melbourne, Australia, 2002.

[Bou75] Boucher J.D. and Ekman P. Facial areas and emotional information. *Journal of Communication*, 25 no. 2:21–29, 1975.

[Bra97] Brainard D. H. The psychophysics toolbox. *Spatial Vision*, 10:433–436, 1997.

[Bur98] Burges J. C. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, pages 121–167, 1998.

[Bus20] Buswell G. T. How people look at pictures. *University of Chicago Press*, 1920.

[Bus04] Busso C., Deng Z., Yildirim S., Bulut M., Lee C.M., Kazemzadeh A., Lee S., Neumann U., Narayanan S. Analysis of emotion recognition using facial expressions, speech and multimodal information. *Proc. 6th International Conference on Multimodal Interfaces (ICMI)*, October State College, PA, 2004.

[Can86] Canny J. A computational approach to edge detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8 no.6:679 – 698, June 1986.

[Cha95] Charbonnier C. *La commande oculaire: étude et validation expérimentale d'interface homme-machine contrôlées par la direction du regard*. PhD thesis, LETI laboratory, Grenoble, France, 1995.

[Che98] Chen L.S., Huang T.S., Miyasato T. and Nakatsu R. Multimodal human emotion/expression recognition. *Proc. International Conference on Face and Gesture Recognition*, pages 396–401, Nara, Japan, 1998.

[Che00] Chen L.S. and Huang T.S. Emotional expressions in audiovisual human computer interaction. *Proc. International Conference on Multimedia and Expo*, pages 423–426, July 30 - August 2 New York City, NY, USA, 2000.

[Cho91] Chou P. Optimal partitioning for classification and regression trees. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13 no. 4:340–354, 1991.

[Coh98] Cohn J.F., Zlochower A. J., Lien J. J and Kanade T. Feature point tracking by optical flow discriminates subtles differences in facial expression. *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 396–401, April Nara, Japan, 1998.

[Coh00] Cohn_Kanade database. http://vasc.ri.cmu.edu/idb/html/face/facial_expression. 2000.

[Coh03a] Cohen I., Cozman F. G., Sebe N., Cirelo M. C. and Huang T. S. Learning bayesian network classifiers for facial expression recognition using both labeled and unlabeled data. *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 16-22 June Madison, Wisconsin, 2003.

[Coh03b] Cohen I., Sebe N., Chen L., Garg A. and Huang T.S. Facial expression recognition from video sequences: temporal and static modelling. *Computer Vision and Image Inderstanding: Special issue on face recognition.*, 91:160–187, July-August 2003.

[Coo98] Cootes T. F., Edwards G. J., Taylor C. J. Active appearance models. *Lecture Notes in Computer Science*, pages 484–491, 1998.

[Dai01] Dailey M., Cottrell G. W. and Reilly J. California facial expressions (cafe). *unpublished digital images, University of California*, San Diego, 2001.

[Dar72] Darwin C. The expression of the emotions in man and animals. London, Murray, 1872.

[De 97] De Silva L.C., Miyasato T. and Nakatsu R. Facial emotion recognition using multimodal information. *Proc. IEEE International Conference on Information, Communication and Signal Processing*, September Singapore, 1997.

[De 00] De Silva L.C. and Ng P.C. Bimodal emotion recognition. *Proc. 4th IEEE International Conference on Automatic Face and Gesture*, pages 332–335, March 28-30 Grenoble, France, 2000.

[Del93] Deliyski D. Acoustic model and evaluation of pathological voice production. *Proc. 3-rd Conference on Speech Communication and Technology EUROSPEECH*, pages 1969–1972, Berlin, Germany, 1993.

[Dem68] Dempster A. A generalization of bayesian inference. *Journal of the Royal Statistical Society, Series B*, pages 205–247, 1968.

[Den04] Deng X., Chang C. H. and Brandle E. A new method for eye extraction from facial image. *Proc. 2nd IEEE international workshop on electronic dessign, test and applications (DELTA)*, 2 no.4:29–34, Perth, Australia, 2004.

[Des49] Descartes H. *Les Passions de l'âme*. Henry Le Gras, Paris, 1649.

[Dev82] Devijver P. R. and Kittler J. *Pattern Recognition: A Statistical Approach*. Prentice-Hall, 1982.

[Din98] Dinges D. F., Mallis M., Maislin G. and Powell J. W. Evaluation of techniques for ocular measurement as an index of fatigue and the basis for alertness management. *Dept. Transp. Highway Safety*, 808 no.762, 1998.

[Don99] Donato G., Barlett M. S., Hager J. C., Ekman P. and Sejnowski T. J. Classifying facial actions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21 no.10:974–989, 1999.

[Dor04a] Dornaika F. and Davoine F. Online appearance-based face and facial feature tracking. *Proc. International Conference on Pattern Recognition ICPR*, pages 814–817, 2004.

[Dor04b] Dornaika F. and dvoine F. Head and facial animation tracking using appearance-adaptive models and particle filters. *Workshop Real-Time Vision for Human-Computer Interaction RTV4HCI in conjunction with CVPR*, 2 July Washington, DC, USA, 2004.

[Dub02] Dubuisson S., Davoine F. and Masson M. A solution for facial expression representation and recognition. *Signal Processing: Image Communication*, 17:657–673, 2002.

[Dud01] Duda R. O., Hart P. E. and Stork D. G. Pattern classification. *John Wiley and Sons*, New York, 2001.

[Ebi93] Ebisawa Y. and Satoh S. Effectiveness of pupil area detection technique using two light sources and image difference method. *Proc. 5th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1268–1269, 1993.

[Ekm72] Ekman P., Friesen W. V. and Ellsworth P. Emotion in the human face. *Pergamon Press*, New York, 1972.

[Ekm78] Ekman P. and Friesen W. V. Facial action coding system. *Consulting Psychologist Press*, 18, no.11:881–905, August 1978.

[Ekm82] Ekman P., Friesen W. V. and Ellsworth P. Research foundations. in p. ekman (ed.) emotion in the human face (2nd ed.). *Cambridge: Cambridge University Press*, 1982.

[Ekm99] Ekman P. The handbook of cognition and emotion: Facial expression. *John Wiley and Sons*, 1999.

[Eng96] Engberg I. S. and Hansen A. V. Documentation on the danish emotional speech database des. *Technical report*, Alborg, September, 1996.

[Ess97] Essa I. A. and Pentland A. P. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 no.7:757–763, 1997.

[Eve01] Eveno N., Caplier A., Coulon P. Y. A new color transformation for lip segmentation. *Proc. IEEE MSSP'01*, September Cannes, France, 2001.

[Eve02] Eveno N., Caplier A., Coulon P. Y. A parametric model for realistic lip segmentation. *Proc. International Conference on Control, Automation, Robotics and Vision (ICARV'02)*, December Singapore, 2002.

[Eve03a] Eveno N. *Segmentation des lèvres par un modèle déformable analytique*. PhD thesis, LIS laboratory, Grenoble, France, 2003.

[Eve03b] Eveno N., Caplier A., Coulon P. Y. Jumping snakes and parametric model for lip segmentation. *Proc. International Conference on Image Processing*, September Barcelone, Espagne 2003.

[Eve04] Eveno N., Caplier A., Coulon P. Y. Automatic and accurate lip tracking. *IEEE Trans. On Circuits and Systems for Video Technology*, 14, no.5:706–715, 2004.

[Eye05] Eyelink Commercial website. http://www.sr-research.com/. 2005.

[Fas03] Fasel B. and Luettin J. Automatic facial expression analysis: A survey. *Pattern Recognition*, 1 no.30:259–275, 2003.

[Fas05] Fasel I., Fortenberry B. and Movellan J. A generative framework for real time object detection and classification. *Computer Vision and Image Understanding*, 98:182–210, 2005.

[Fer03] Feret database. http://www.itl.nist.gov/iad/humanid/feret/feret_master.html. 2003.

[Gao99] Gao Y. and Leung M.K.H. Human face recognition using line edge maps. *Proc. IEEE 2nd Workshop Automatic Identification Advanced Technology*, pages 173–176, October 1999.

[Gao03] Gao Y., Leung M.K.H, Hui S.C. and Tananda M.W. Facial expression recognition from line-based caricatures. *IEEE Transaction on System, Man and Cybernetics - PART A: System and Humans*, 33 no.3, May 2003.

[Gou00] Gouta K. and Miyamoto M. Facial areas and emotional information. *Japanese journal of psychology*, 71 no. 3:211–218, 2000.

[Ham03] Hammal_Caplier database. http://facialexpressions.free.fr/. 2003.

[Han81] Hand D.J. Discrimination and classification. *John Wiley and Sons*, 1981.

[Han05] Hansen D. W., Pece A. E. C. Eye tracking in the wild. *Computer Vision and Image Understanding*, 98:155–181, 2005.

[Hje01] Hjelmäs H. and Low B. Face detection: a survey. *Proc. Computer Vision and Understanding*, 83:236–274, 2001.

[Hér96] Hérault J. A model of colour processing in the retina of vertebrates: from photoreceptors to colour opposition and colour constancy. *Neurocomputing*, 12:113–129, 1996.

[Hua97] Huang C.L. and Huang Y.M. Facial expression recognition using model-based feature extraction and action parameters classification. *Visual Communication and Image Representation*, 8 no. 3:278–290, 1997.

[Hut89] Hutchinson Jr. T., Reichert K., Frey L. Human-computer interaction using eye-gaze input. *IEEE Transaction System Man Cybernet*, 19:1527–1533, 1989.

[INT00] INTERFACE Project. Multimodal analysis/synthesis system for human interaction to virtual and augmented environments. *EC IST-1999-No 10036, coo F. Lavagetto, http://www.ist-interface.org*, 2000.

[Ji 01] Ji Q. and Yang X. Real time visual cues extraction for monitoring driver vigilance. *Lecture Notes Computer Science*, 2095, 2001.

[Ji 02] Ji Q. and Zhu Z. Eye and gaze tracking for interactive graphic display. *Proc. the 2nd International Symposium on Smart Graphics*, pages 79–85, 2002.

[Ji 04] Ji Q., Zhu Z. and Lan P. Real-time non intrusive monitoring and prediction of driver fatigue. *IEEE Transactions on Vehicular Technology*, 53 no.4:1052–1068, July 2004.

[Jus01] Juslin P.N. and Laukka P. Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion*, 1 no.4:381–412, 2001.

[Jus03] Juslin P. N. and Laukka P. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, pages 770–814, 2003.

[Kal02] Kallel R., Cottrell M. and Vigneron V. Bootstrap for neural model. *Neurocomputing*, 48:175–183, 2002.

[Kap98] Kapmann M. and Zhang L. Estimation of eye, eyebrow and nose features in videophone sequences. *International Workshop on Very Low Bitrate Video Coding (VLBV 98)*, pages 101–104, October Urbana, USA, 1998.

[Kas88] Kass M., Withins A. and D. Tersopolos. Snakes : Actives contours models. *International Journal of computer vision*, 1 no.4:321–331, January 1988.

[Kas01] Kashima H., Hongo H., Kato K. and Yamamoto K. A robust iris detection method of facial and eye movement. *Proc. Vision Interface Annual Conference*, 7-9 June Ottawa, Canada, 2001.

[Kje96] Kjedlsen R. and Kender J. Finding skin in color images. *Face and gesture Recognition*, pages 312–317, 1996.

[Ko 99] Ko J-G, Kim K-N and Ramakrishma R. S. Facial feature tracking for eye-head controlled human computer interface. *IEEE TENCON*, September Cheju, Korea, 1999.

[Kon94] Kononenko I. Estimating attributes: Analysis and extension of relief. *Proc. European conference on Machine Learning*, pages 171–182, 1994.

[Lee04] Lee C.M, Yuildrim S., Bulut M., Kazemzadeh A., Busso C., Deng Z., Lee S., Narayanan S. Emotion recognition based on phoneme classes. *Proc. International Conference on Spoken Language Processing*, Jeju Island (Corea), 2004.

[Lei83] Leigh J. R. and Zee D. S. The neurology of eye movements. *FA Davis Company*, Philadelphia, 1983.

[Lie98] Lien J.J, Kanade T., Cohn J.F. and Li C. Subtly different facial expression recognition and expression intensity estimation. *Proc. Computer Vision and Pattern Recognition (CVPR)*, pages 853–859, June 23-25 Santa Barbara, CA, 1998.

[Lin80] Linde Y., Buzo A. and Gray R. M. An algorithm for vector quantizer design. *IEEE Trans. on Communication*, 28, no.1:84–95, January 1980.

[Luc81] Lucas B. D. and Kanade T. An iterative image registration technique with an application to stereo vision. *Proc. of International Joint Conference on Artificial Intelligence*, 18, no.11:674–680, August Vancouver, Canada, 1981.

[Lyo98] Lyons M.J. and Akamatsu S. Coding facial expressions with gabor wavelets. *Proc. Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205, 14-16 April Nara, Japan, 1998.

[Mac99] Machine Perception Toolbox. Face detection algorithm: http://mplab.ucsd.edu/grants/project1/freesoftwar/mptwebsite/introductionframe.html. *Proc. Of the SPIE: On Storage and Retrieval for Images and video Databases*, 1, no.3656:458–466, 1999.

[Mac06] Mac Graw-Hill. http://duskin.com/connectext/psy/ch10/facex.mhtml. 2006.

[Mai00] Maio D. and Maltoni D. Real-time face location on grayscale static images. *Pattern Recognition*, 33:1525–1539, 2000.

[Mal01] Malciu M. and Preteux F. Mpeg-4 compliant tracking of facial features in video sequences. *Proc. of International Conference on Augmented, Virtual Environments and 3D Imaging*, pages 108–111, May Mykonos, Greece, 2001.

[Mas98] Massaro D.W. Illusions and issues in bimodal speech perception. *Proc. Auditory Visual Speech Perception*, pages 21–26, December Terrigal-Sydney Australia, 1998.

[Mat00] Matsumoto Y., Ogasawara T. and Zelinski A. Behavior recognition based on head and gaze direction measurement. *Proc. IEEE International Conference on Intelligent Robots and systems (IROS)*, pages 2127–2132, Takamatsu, Japan, 2000.

[Med91] Medan Y., Yair E. and Chazan D. Super resolution pitch determination of speech signals. *Trans. IEEE Signal Processing*, 39:40–48, January 1991.

[Meh68] Mehrabian A. Communication without words. *Psychology Today*, 2 no.4:53–56, 1968.

[Mor99] Morimoto C., Koons D., Flickner M. and Zhai S. Keeping an eye for hci. *Proc. 12th Brazilian Symp on Computer Graphics and Image Processing*, pages 171–176, October Campinas, Brazil 1999.

[Nog01] Nogueiras A., Moreno A., Bonafonte A. and Marino J.B. Speech emotion recognition using hidden markov models. *Proc. European Conference on Speech Communication and Technology*, Scandinavia, 2001.

[Nou05] Noureddin B., Lawrence P. D., Man C. F. A non-contact for tracking gaze in a human-computer interface. *Computer Vision and Image Understanding*, 98:52–82, 2005.

[Oli00] Oliver N., Pentland A. and Bérard F. Lafter: a real-time face and tracker with facial expression recognition. *Pattern Recognition*, 33:1369–1382, 2000.

[Ope06] OpenCv. http://www.intel.com/technology/computing/opencv/index.htm. 2006.

[Orl94] Orl database. http://www.cl.cam.ac.uk/reaserch/dtg/attarchive/facedatabase.html. 1994.

[Ots78] Otsu N. A threshold selection method from gray level histogram. *IEEE Transaction on Systems, Man, and Cybernetics*, 8:62–66, 1978.

[Pan00a] Pantic M. and Rothkrantz L. J. M. Expert system for automatic analysis of facial expressions. *Image and Vision Computing Journal*, 18, no.11:881–905, August 2000.

[Pan00b] Pantic M. and Rothkrantz L.J.M. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 22 no.12:1424–1445, December 2000.

[Pan03] Pantic M. and Rothkrantz L.J.M. Toward an affect-sensitive multimodal human-computer interaction. *Proc. Procedings of the IEEE*, 91 no. 9:1370–1390, September 2003.

[Pan05a] Pantic M. and Patras I. Detecting facial actions and their temporal segmentation in nearly frontal-view face image sequences. *Proc. IEEE International Conference on Systemsn Man and Cybernetics*, October Waikoloa, Hawaii 2005.

[Pan05b] Pantic M., Valstar M.F., Rademaker R. and Maat L. Web-based facial expression database. *Proc. IEEE International Conference on Multimedia and Expo. (www.mmifacedb.com)*, pages 317–321, July Amsterdam, The Netherlands, 2005.

[Pan06] Pantic M. and Patras I. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transaction on System Systemsn, Man, and Cybernetics_Part B: CYBERNETICS*, 36 no.2, April 2006.

[Par99] Pardas M. Automatic face analysis for model calibration. *Proc. International Workshop on Synthetic Natural Hybrid Coding and three Dimentional Imaging (IWSNHC3DI)*, pages 12–15, September 1999.

[Par00] Pardas M. Extraction and tracking of the eyelids. *Proc. International Conference on Acoustics, Speech and Signal Processing ICASSP*, 4:2357–2360, June Istambul, Turkey, 2000.

[Par01] Pardas M., Sayrol E. Motion estimation based tracking of active contours. *Pattern recognition letters*, 22 no.13:1447–1456, November, 2001.

[Par02] Pardas M., Bonafonte A. Facial animation parameters extraction and expression detection using hmm. *Signal Processing: Image Communication*, 17:675–688, 2002.

[Pel97] Pelli D. G. The video toolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10 no.4, 1997.

[Pen00] Pentland A. Looking at people. *Communications of the ACM*, 43 no.3:35–44, 2000.

[Pet00] Petrushin V. A. Emotion recognition in speech signal: experimental study, development, and application. *Proc. 6th International Conference on Spoken Language Processing*, Beijing, China, 2000.

[Prê92] Prêteux F. *On a distance function approach for grey-level mathematical morphology.* E. R. Dougherty Eds., Mathematical Morphology in Image Processing, Dekker M., 1992.

[Qia03] Qiang J. and Zhu Z. Non-intrusive eye and gaze tracking for natural human computer interaction. *Journal. MMI interaktiv*, 6, 2003.

[Qua01] Quatieri T. F. Discrete time speech signal processing: Principles and practice. 2001.

[Rab89] Rabiner L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77, no.2:257–286, February 1989.

[Rad93] Radeva P. and Serrat J. Rubber snake: Implementation on signed distance potential. *Proc. Proceedings SWISS VISSION*, pages 187–194, September 1993.

[Rad95] Radeva P. and Marti E. Facial features segmentation by model-based snakes. *Proc. International Conference on Computer Analysis and Image Processing*, pages 515–520, September 6-8 Prague, 1995.

[Ros96] Rosenblum M., Yacoob Y. and Davis L.S. Human expression recognition from motion using a radial basis function network architecture. *IEEE Trans. Neural Networks*, 7 no.5:1121–1137, 1996.

[Row00] Roweis S. and Saul L. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 no.5500:2323–2326, 2000.

[Sat01] Sato H., Mitsukura Y, Fukumi M., Akamatsu N. Emotional speech classification with prosodic parameters by using neural networks. *Proc. Australien and New Zealand Intelligent Information Systems Conference*, pages 395–398, 2001.

[Sch03a] Scherer K. R. Vocal communication of emotion. *A review of research paradigms. Speech Communication*, 40:227–256, 2003.

[Sch03b] Schröder M. *Speech and Emotion Research.* PhD thesis, 2003.

[Seb04] Sebe N., Cohen I., Huang T.S. Multimodal emotion recognition. Jun, 18 2004.

[Sha76] Shafer G. *A Mathematical Theory of Evidence.* Princeton University Press, 1976.

[Shi94] Shi J. and Tomasi C. Good features to track. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, Seattle, Washington, 1994.

[Sib01] Sibert L. E., Templeman J. N. and Jacob R.J.K. Evaluation and analysis of eye gaze interaction. *NRL Report NRL/FR/5513 01-9990*, December Washington, 2001.

[Sim04] SimilarNet the European Taskforce for Creating Human-Machine Interfaces Similar to Human-Human Communication,. http://www.similar.cc/. 2004.

[Sme90] Smets P. The combination of evidence in the transferable belief model. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12:447–458, 1990.

[Sme94] Smets P. and Kruse R. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.

[Sme98] Smets P. Handbook of defeasible reasoning and uncertainty manage-ment system: The transferable belief model for quantified belief representation. *Kluwer*, 1:267–301, 1998.

[Sme00] Smets P. Data fusion in the transferable belief model. *Proc. of International Conference on Information Fusion*, pages 21–33, July Paris, France, 2000.

[Smi03] Smith P., Shah M., and Lobo N.V. Determining driver visual attention with one camera. *IEEE Transaction on Intelligent Transportation Systems*, 4 no.4, December 2003.

[Smi05] Smith M. L., Cottrell G. W., Gosselin F., and Schyns P. G. Transmitting and decoding facial expressions. *Psychological Science*, 8, no.6:679–698, March 2005.

[Sou03] Sourd A. *Master 2 report, Laboratoire de Psychologie Sociale, Grenoble*, 2003.

[Tal99] Talmi K. and Liu J. Eye and gaze tracking for visually controlled interactive stereoscopic displays. *Signal Processing: Image Communication*, 14:799–810, 1999.

[Tan02] Tan K., Kriegman D. and Ahuja N. Appearance-based eye gaze estimation. *Proc. Workshop on Applications of Computer Vision*, pages 191–195, 2002.

[Tao98] Tao H. and Huang T.S. Connected vibration: A model analysis approach to non-rigid motion tracking. *Proc. IEEE Computer Vision and Pattern Recognition*, pages 735–740, 1998.

[Tao99] Tao H. and Huang T.S. explanation-based facial motion tracking using a piecewise bezier volume deformation model. *Proc. International Conference on Computer Vision and Pattern Recognition*, 1:611–617, 1999.

[Tek99] Tekalp M. Face and 2d mesh animation in mpeg-4. *Tutorial Issue on the MPEG-4 Standard, Image Communication Journal, Elsevier*, 1999.

[Tia00] Tian Y., Kanade T. and Cohn J. Dual state parametric eye tracking. *Proc. 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 110–115, March Grenoble, 2000.

[Tia01] Tian Y., Kanade T. and Cohn J.F. Recognizing action units for facial expression analysis. *Trans. IEEE Pattern Analysis and Machine Intelligence*, 23 no.2:97–115, February 2001.

[Tor99] Torralba A.B. and Hérault J. An efficient neuromorphic analog network form motion estimation. *IEEE Trans. on Circuits and Systems-I: Special Issue on Bio-Inspired Processors and CNNs for Vision*, 46 no.2, 1999.

[Tsa00] Tsapatsoulis N., Karpouzis K., Stamou G., Piat F. and Kollias S. A fuzzy system for emotion classification based on the mpeg-4 facial definition parameter set. *Proc. 10th European Signal Processing Conference*, September 5-8 Tampere, Finland, 2000.

[Tse98] Tsekeridou S. and Pitas I. Facial feature extraction in frontal views using biometric analogies. *Proc. 9th European Signal Processing Conference*, 1:315–318, September 8-11 Island of Rhodes, Greece, 1998.

[Val00] Valet L., Mauris G. and Bolon P. A statistic overview of a recent literature in information fusion. *Proc. International conference in information fusion*, pages MoC3 22–29, Paris, France, 2000.

[Ver04] Ververidis D., Kotropolos C. Automatic speech classification to five emotional states based on gender information. *Proc. 12th European Signal Processing Conference*, pages 341–344, Vienna, 2004.

[Vez03] Vezhnevets V. and Degtiareva A. Robust and accurate eye contour extraction. *Proc. GraphicsCon*, pages 81–84, September 5-1O Moscow, Russia, 2003.

[Wan05] Wang J.G., Sung E. and Venkateswarlu R. Estimating the eye gaze from one eye. *Computer Vision and Image Understanding*, 98:83–103, 2005.

[Yac96] Yacoob Y. and Davis L.S. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18 no.6:636–642, June 1996.

[Yag94] Yager R.R, Fedrizzi M. and Kacprzyk J. Advance in the dempster-shafer theory of evidence. *John Wiley & Sons: New York,*, 1994.

[Yal97] Yale database. http://cvc.yale.edu/projects/yalefaces/yalefaces.html. 1997.

[Yan99] Yang M. H. and Ahuja N. Gaussian mixture model for human skin color and its application in image and video database. *Proc. Of the SPIE: Conf. On Storage and Retrieval for Images and video Databases*, 1, no.3656:458–466, San Jose, CA, USA, 1999.

[Yoo05] Yoo D. H. and Chang M. J. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *Computer Vision and Image Understanding*, 98:25–51, 2005.

[Yos00] Yoshitomi Y., Kim S., Kawano T. and Kitazoe T. Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. *Proc. ROMAN*, pages 178–183, 2000.

[Yui92] Yuille A., Hallinan P. and Cohen D. Feature extraction from faces using deformable templates. *International Journal of computer Vision*, 8 no.2:99–111, August 1992.

[Zen05] Zeng Z., Tu Jilin, Pianfetti B., Liu M., Zhang T., Zhang Z., Huang T.S. and Levinson S. Audio-visual affect recognition through multi-stream fused hmm for hci. *Proc. International Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, San Diego, CA, USA 2005.

[Zha96] Zhang L. Estimation of eye and mouth corner point positions in a knowledge based coding system. *Proc. SPIE*, 2952:21–28, Digital Compression Technologies and Systems for Video Communications, October, 1996.

[Zha97] Zhang L. Tracking a face for knowledge based coding of videophone sequences. *Signal Processing: Image Communication*, 10 no.13:93–114, July 1997.

[Zha98] Zhang Z., Lyons L., Schuster M. and Akamatsu S. Comparison between geometry-based and gabor wavelets-based facial expression recognition using multi-layer perceptron. *Proc. Automatic Face and Gesture Recognition*, pages 454–459, Japan, 1998.

[Zha05] Zhang Y. and Qiang J. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27 no.5:699–714, May 2005.

[Zhu05] Zhu Z. and Qiang J. Robust real-time eye detection and tracking and variable lighting conditions and various face conditions. *Computer Vision and Image Understanding*, 98:124–154, 2005.

Publications

[Buc06] Buciu I., Hammal Z., Caplier A., Nikolaidis N. and Pitas I. Enhencing facial expression classification by information fusion. *Proc. 14th European Signal Processing Conference (EUSIPCO)*, 2006.

[Ham03] Hammal Z., Eveno N., Caplier A. and Coulon P.Y. Extraction réaliste des traits caractéristiques du visage à l'aide de modèles paramétriques adaptés. *Proc. 19ème colloque sur le traitemant du signal et des images (Gretsi)*, Paris, 2003.

[Ham04a] Hammal Z., Caplier A. Eyes and eyebrows parametric models for automatic segmentation. *Proc. Southhwest Symposium on Image Analysis and Interpretation. (IEEE SSIAI)*, Nevada, USA, 2004.

[Ham04b] Hammal Z., Caplier A. Analyse dynamique des transformations des traits du visage lors de la production d'une emotion. *Proc. Atelier sur l'analyse du geste (RFIA)*, Toulouse, 2004.

[Ham04c] Hammal Z., Caplier A. and Rombaut M. Classification d'expressions faciales par la theorie de l'évidence. *Rencontre Francophones sur la Logique Floue et ses Applications (LFA)*, Nantes, 2004.

[Ham05a] Hammal Z., Bozkurt B., Couvreur L., Unay D., Caplier A. and Dutoit T. Classification bimodale d'expressions vocales. *20ème colloque sur le traitemant du signal et des images (Gretsi)*, Louvain-la-Neuve, Belgique, 2005.

[Ham05b] Hammal Z., Bozkurt B., Couvreur L., Unay D., Caplier A. and Dutoit T. Passive versus active: Vocal classification system. *Proc. 13th European Signal Processing Conference, (EUSIPCO)*, Turkey, 2005.

[Ham05c] Hammal Z., Caplier A. and Rombaut M. Belief theory applied to facial expressions classification. *Proc. 3rd International Conference on Advances in Pattern Recognition (ICAPR)*, Bath, United Kingdom, 2005.

[Ham05d] Hammal Z., Caplier A. and Rombaut M. A fusion process based on belief theory for classification of facial basic emotions. *Proc. the 8th International Conference on Information fusion, (ISIF)*, Philadelphia, PA, USA, 2005.

[Ham05e] Hammal Z., Couvreur L., Caplier A. and Rombaut M. Facial expressions recognition based on the belief theory: Comparison with diferent classifiers. *Proc. 13th International Conference on Image Analysis and Processing. (ICIAP)*, Italy, 2005.

[Ham05f] Hammal Z., Eveno N., Caplier A. and Coulon P.Y. Extraction des traits caracteristiques du visage à l'aide de modèle paramétriques adaptés. *Traitement du signal (TS)*, 22 no.1:59–71, 2005.

[Ham05g] Hammal Z., Massot C., Bedoya G. and Caplier A. Eyes segmentation applied to gaze direction and vigilance estimation. *Proc. 3rd International Conference on Advances in Pattern Recognition (ICAPR)*, Bath, United Kingdom, 2005.

[Ham06a] Hammal Z. Dynamic facial expression understanding based on temporal modeling of transferable belief model. *Proc. International conference on computer vision theory and application (VISAPP)*, Setubal, Portugal, 2006.

[Ham06b] Hammal Z., Caplier A. and Rombaut M. Fusion for classification of facial expressions by the belief theory. *Journal of advances in information fusion (JAIF)*, (Submitted) 2006.

[Ham06c] Hammal Z., Couvreur L., Caplier A. and Rombaut M. Facial expressions classification: A new approach based on transferable belief model. *International Journal of Approximate Reasoning (Elsevier)*, (Under revision) 2006.

[Ham06d] Hammal Z., Eveno N., Caplier A. and Coulon P.Y. Parametric models for facial features segmentation. *Signal processing*, 86:399–413, 2006.

Facial features segmentation, analysis and recognition of facial expressions by the Transferable Belief Model

The aim of this work is the analysis and the classification of facial expressions. Experiments in psychology show that human is able to recognize the emotions based on the visualization of the temporal evolution of some characteristic fiducial points. Thus we firstly propose an automatic system for the extraction of the permanent facial features (eyes, eyebrows and lips). In this work we are interested in the problem of the segmentation of the eyes and the eyebrows. The segmentation of lips contours is based on a previous work developed in the laboratory. The proposed algorithm for eyes and eyebrows contours segmentation consists of three steps: firstly, the definition of parametric models to fit as accurate as possible the contour of each feature; then, a whole set of characteristic points is detected to initialize the selected models in the face; finally, the initial models are finally fitted by taking into account the luminance gradient information. The segmentation of the eyes, eyebrows and lips contours leads to what we call *skeletons of expressions*. To measure the characteristic features deformation, five characteristic distances are defined on these skeletons. Based on the state of these distances a whole set of logical rules is defined for each one of the considered expression: *Smile*, *Surprise*, *Disgust*, *Anger*, *Fear*, *Sadness* and *Neutral*. These rules are compatible with the standard MPEG-4 which provides a description of the deformations undergone by each facial feature during the production of the six universal facial expressions. However the human behavior is not binary, a pure expression is rarely produced. To be able to model the doubt between several expressions and to model the unknown expressions, the Transferable Belief Model is used as a fusion process for the facial expressions classification. The classification system takes into account the evolution of the facial features deformation in the course of the time. Towards an audio-visual system for emotional expressions classification, a preliminary study on vocal expressions is also proposed.

Keywords: segmentation, iris, eyes, eyebrows, facial expressions, Transferable Belief Model, vocal expressions.

Laboratoire des Images et des Signaux
46 avenue Felix Viallet
38031 Grenoble Cedex

Segmentation des Traits du Visage, Analyse et Reconnaissance d'Expressions Faciales par le Modèle de Croyance Transférable

L'objectif de ce travail est l'analyse et la classification d'expressions faciales. Des expériences en psychologie ont permis de mettre en évidence le fait que l'être humain est capable de reconnaître les émotions sur un visage à partir de la visualisation de l'évolution temporelle de certains points caractéristiques de celui-ci. Nous avons donc tout d'abord proposé un système d'extraction automatique des contours des traits permanents du visage (yeux, sourcils et lèvres). Dans ce travail nous nous intéressons au problème de la segmentation des yeux et des sourcils. La segmentation des contours des lèvres est basée sur un travail précédent développé au sein du laboratoire. L'algorithme proposé pour l'extraction des contours des yeux et des sourcils est constitué de trois étapes : d'abord la définition de modèles paramétrique pour modéliser au mieux le contour de chaque trait; ensuite, les modèles choisis sont initialisés sur les images à segmenter grâce à l'extraction d'un ensemble de points caractéristiques; enfin, les modèles initiaux sont ajustés finement en tenant compte d'information de gradient de luminance. La segmentation des contours des yeux, des sourcils et des lèvres conduit à ce que nous appelons des *squelettes d'expressions*. Pour mesurer la déformation des traits caractéristiques, cinq distances caractéristiques sont définies sur ces squelettes. Basé sur l'état de ces distances un ensemble de règles logiques est défini pour chacune des expressions considérées: *Sourire*, *Surprise*, *Dégoût*, *Colère*, *Peur*, *Tristesse*, *Neutre*. Ces règles sont compatibles avec la norme MPEG-4 qui fournit une description des transformations subies par chacun des traits du visage lors de la production des six expressions faciales universelles. Cependant le comportement humain n'étant pas binaire, une expression pure est rarement produite. Pour pouvoir modéliser le doute entre plusieurs expressions et le cas des expressions inconnues, le Modèle de Croyance Transférable est utilisé comme processus de fusion pour la classification des expressions faciales. Le system de reconnaissance développé tient compte de l'évolution au cours du temps des déformations des traits du visage. Dans la perspective d'un système audio-visuel de reconnaissance d'expressions émotionelles, une étude préliminaire sur des expressions vocales a aussi été menée.

Mots-clefs: segmentation, iris, yeux, sourcils, expressions faciales, Modèle de Croyance Transferable, expressions vocales.

Laboratoire des Images et des Signaux
46 avenue Felix Viallet
38031 Grenoble Cedex